CONVEGNO
SIpEIA
CONFERENCE
2026

BOOK OF ABSTRACTS

Facoltà di Lettere e Filosofia
Sapienza Università di Roma
Rome, 2-3 February 2026

INDICE
TABLE OF CONTENTS

KEYNOTE SPEECHES

Sanmay Das
Virginia Tech

On AI Alignment in the Provision of Social Services: Opportunities and Challenges

Artificial intelligence is increasingly used to aid decision-making about the allocation of scarce societal resources, for example housing for homeless people, organs for transplantation, and educational supports for students. What does it mean for these systems to be "aligned" with human preferences? In practice, this involves attempts to achieve some combination of fairness, efficiency, and incentive compatibility, depending on the preferences of some set of stakeholders. In this talk I will discuss how the theories of local justice and of street level bureaucracy can inform our use of AI in high-stakes decision making about social services. I will focus mostly on prioritization of services for those experiencing homelessness, although the ideas are broadly applicable. I will give a peak into theoretical, empirical, and experimental results, and discuss where I think AI can be most helpful, as well as significant human and technical challenges.

***

Francesca Rossi
IBM Fellow and Global Leader for Responsible AI and AI Governance

AI ethics and governance: an evolving landscape with a clear ROI

Supported by AI, we will be able to make more grounded decisions and to focus on the main values and goals of a decision process rather than on routine and repetitive tasks. However, such a powerful technology also raises some concerns, related to privacy, fairness, value alignment, explainability, accountability, transparency, misinformation, deep fakes, copyright, and much more.

These concerns are among the obstacles that hold AI back or that cause worry for current AI users, adopters, and policy makers. Without concrete answers to these questions, many will not trust AI, and therefore will not fully adopt it nor get its positive impact.

In this talk I will present the main issues around AI ethics and safety and how they have evolved over the years, as AI capabilities advanced. I will also discuss the ROI of AI ethics, discussing both instrumental and value-based approaches to investing in AI ethics and governance.

***

Daniel Innerarity
AI and Democracy Chair – EUI

People in AI

Digital technologies seem to be technologies that avoid people. However, for AI to be truly democratic, people should be present along its life cycle in different moments, different ways and according to what is at stake.

Mariarosaria Taddeo
Oxford Internet Institute- University of Oxford

The Ethics of Artificial Intelligence in Defence

Why we need an ethics of AI in defence. Over the past two decades, there have been growing efforts to design, develop and deploy digital technologies for national defence. Defence agencies across the globe identify AI as a key technology to maintain an edge over adversaries. As a result, efforts to develop or acquire AI capabilities for defence are growing on a global scale. Unfortunately, they remain unmatched by efforts to define ethical frameworks to guide the use of AI in the defence domain. The question is whether and how we can leverage AI capabilities in defence while ensuring that they remain aligned to the fundamental values of our societies.

***

Vincenzo Paglia
President Emeritus of the Pontifical Academy for Life

L'umano nell'era dell'IA: responsabilità, dignità, bene comune

L'intelligenza artificiale non è solo un insieme di strumenti potenti: segna un cambio di paradigma che tocca lavoro, cura, educazione, informazione, politica e persino la comprensione della persona. Per questo serve un discernimento etico capace di orientare innovazione e governance al bene comune. Occorre riconoscere con chiarezza la differenza tra umano e macchina: l'IA elabora e correla dati, ma non è un soggetto morale; non possiede coscienza, esperienza, libertà né responsabilità. Il rischio decisivo è il paradigma tecnocratico che riduce l'umano a dato, profilo, prestazione, e la delega opaca di decisioni ad alto impatto (sanità, servizi, giustizia, sicurezza), con conseguenze su diritti, disuguaglianze e fiducia pubblica, anche di fronte a disinformazione e manipolazione. Da qui l'urgenza di "umanizzare la tecnica": trasparenza, inclusione, responsabilità/accountability, imparzialità, affidabilità, sicurezza e tutela della privacy come criteri guida, nella scia della Rome Call. L'orizzonte è un nuovo umanesimo, capace di integrare scienza, diritto, politica ed etica, promuovendo cooperazione e regole condivise affinché l'IA serva la dignità di ogni persona e la casa comune.

1.
EPISTEMOLOGIA E AFFIDABILITÀ
EPISTEMOLOGY AND TRUSTWORTHINESS

Contributors:

1. Guido Boella
2. Arianna Boldi, Ilaria Gabbatore, Francesca Marina Bosco
3. Tomasz Braun
4. Valentina Cavani
5. Carlo Ciucani, Eugenia Polizzi, Giulia Andrighetto
6. Demet Tugce Dumanoglu Cosgrave
7. Antonio Pio De Mattia
8. Eylem Doğan
9. Mustapha El Moussaoui
10. Indrė Espinoza
11. Antonio Gaitán-Torres
12. Michael Hemmingsen
13. Chenithung L. Ngullie
14. Gabriele Nino, Francesca Alessandra Lisi
15. Marica Notte, Vieri Giuliano Santucci
16. Teresa Numerico
17. Luca Pezzini
18. Latha Poonamallee
19. Fathia Sabiella
20. Luigi Scorzato
21. Ludovica Schaerf
22. Cristina Voto
23. Xiaotong Li

Guido Boella
Computer Science Department, Università degli studi di Torino

Towards an Ethics Sensitive to Particulars: How Deep Learning Goes Beyond Universals through Tacit Knowledge

Throughout history, every technological revolution has served not only as a tool but also as a metaphor for understanding the human mind. From the mechanical automata of Descartes to the symbolic computers of the twentieth century, each era has projected its dominant technology onto models of cognition. Deep learning represents the latest—and perhaps most transformative—iteration of this pattern. Unlike symbolic AI, which sought to reproduce intelligence through explicit rules and universal principles, deep neural networks learn autonomously from data, embodying a form of implicit, contextualized understanding that closely parallels Michael Polanyi's notion of tacit knowledge. Tacit knowledge refers to forms of knowing that are embodied, situational, and difficult to formalize—skills and intuitions that cannot be fully articulated. Similarly, deep learning systems operate as "black boxes" capable of capturing subtle regularities in complex, multidimensional data without explicit instruction. Their success across chaotic and nonlinear domains such as meteorology, biology, and language suggests that cognition — and ethical reasoning and law — may rely less on universal rules and more on sensitivity to particulars. This shift from explicit

representation to emergent learning can mirror a broader philosophical transition from universalistic ethics, typified by Kantian deontology and utilitarianism, to relational frameworks such as the Ethics of Care developed by Carol Gilligan. As argued by David Weinberger, deep learning's capacity to model complexity without reducing it to simple generalizations offers a metaphor for moral reasoning attentive to context, empathy, and interdependence. While deep learning obviously lacks genuine empathy (but it can increasingly simulate it, as it appears from LLMs), its architecture exemplifies a paradigm that values particularity and complexity over abstraction and uniformity.

Deep learning, by modeling knowledge through adaptive, context-sensitive processes rather than fixed universal rules, can also provide a new lens for understanding law. The Western ideal of universality and objectivity in law—rooted in Roman and Enlightenment rationalism—is itself a cultural construct. Legal systems in Arabic and Eastern traditions have long emphasized context, relationships, and moral reciprocity over abstract universality. In this light, deep learning's attention to particulars and situational complexity resonates with non-Western legal epistemologies, suggesting that both AI and jurisprudence may evolve beyond the rigid universalism of modernity toward a more pluralistic, relational, and context-aware model of justice thus, deep learning can be interpreted not merely as a computational method but as a model of tacit, situated intelligence that challenges the dominance of universalistic ethics and law in both moral philosophy and legal studies, and in AI design—inviting a more relational, context-aware approach to artificial and human understanding alike.

Keywords: Ethics and AI; Law; Deep Learning; Tacit Knowledge

Guido Boella is a Professor of Computer Science at the University of Turin where he has been Vice-Rector for AI. He co-founded SIpEIA, the Italian Society for the Ethics of Artificial Intelligence, and is co-coordinator of the Magazine Intelligenza Artificiale magia.news. His research connects AI, law, and ethics, exploring how intelligent systems interact with social and normative frameworks. Boella promotes responsible and transparent AI through academic, industrial, and civic initiatives such as the CIM Competence Center, he is vice president of, and civic technologies he developed, like the social network FirstLife and the blockchain app CommonsHood. He is the coordinator of several EU projects such as CO3, NLAB4Citizens, DUT CORPUS, CHEDIH and PAI EDIH.

Arianna Boldi
Department of Psychology – GIPSI Research Group, Università degli studi di Torino

Ilaria Gabbatore
Department of Humanities – GIPSI Research Group, Università degli studi di Torino

Francesca Marina Bosco
Department of Psychology – GIPSI Research Group – Neuroscience Institute of Turin – NIT, Università degli studi di Torino

Pragmatic Norms, Machine Judgment, and Clinical Stakes: Probing LLM "Understanding" with a Validated Protocol

What, if anything, does an LLM "understand" when it assigns a normative judgment to a human utterance? We address this question by using a clinically validated framework (ABaCo, Assessment Battery for Communication) to treat a state-of-the-art LLM as an autonomous coder of pragmatic ability, i.e., the abil-

ity to appropriately use language in a given context. The approach is philosophically motivated: ABaCo operationalizes inferential and contextual aspects of meaning (e.g., indirect speech acts, irony, deceit, contextual appropriateness) and thereby provides a concrete testbed for the current debate about the reliability and trustworthiness of AI judgment. Empirically, pilot evidence shows that the model tends to align with expert coders on straightforward communicative acts, while revealing systematic divergences on items requiring higher-order inference or sensitivity to implicatures and norm violations. We analyze what these patterned disagreements tell us about the locus of error: literalist biases, over-ascription of understanding, and rigidity in applying criteria. These error profiles have immediate ethical implications for human-centric deployment. In particular, when LLM outputs are used to scaffold clinical or educational workflows, reliability must be understood as domain- and phenomenon-relative, with mandatory human oversight precisely where pragmatic opacity is highest. The contribution thus links what the model can warrant to where and how it may be safely used. Moreover, it offers a reproducible template, combining inter-rater agreement with qualitative inspection, for evaluating claims about AI "understanding" in contexts where accurate normative assessment matters.

Keywords: Pragmatics; Large Language Models; Clinical Assessment; Digital Health; Human-AI collaboration

Arianna Boldi is a post-doctoral researcher in Psychology at the University of Turin and a licensed psychologist. She got her Ph.D. in psychology at the University of Turin, while working at the intersection between Psychology and Human-Computer Interaction. Her current research work examines human–AI interaction, with specific regard to pragmatic communication in conversational agents. Since 2020, she has published in several HCI and psychology-related journals (HCI, IJHCI, IJHCS, TOCHI, BIT) and established international collaborations on the topic of behavior-change technologies, video games, and conversational agents.

Ilaria Gabbatore is a psychologist and a psychotherapist. She got her Ph.D. in neuroscience- Cognitive Science at the University of Turin. After working for a few years at the University of Oulu (Finland), she moved back to Italy where is currently a tenure-track assistant professor (rtdB) at the Department of Humanities, University of Turin. Her research interests lie in the field of pragmatic communication, from both a developmental and a neuropsychological perspective, with a particular interest in the assessment and treatment of cognitive and communicative impairment in aging and clinical conditions.

Francesca Marina Bosco is a cognitive psychologist and cognitive psychotherapist, full professor in General psychology and Psychology of communication and communication disorders at the Department of Psychology, University of Turin, Italy. She is coordinator of the GIPSI – research Group on Inferential Processes in Social Interaction at the University of Turin. Her research interests concern the cognitive processes underlying pragmatic communication in healthy and pathological conditions, with a particular attention to developmental aspects and to the rehabilitation and enhancement of communicative abilities.

Tomasz Braun
Lazarski University

The Collective Agency of Artificial Intelligence: Autonomy without Personhood

The ability of AI to interpret data evokes a multitude of philosophical, societal and legal questions. Despite of a variety of academic debates they are hardly responded so far. Some of those questions refer to autonomy of AI and its similarity to collective agency understood as a capacity of a group to act jointly and independently.

Machines have ability to forecast the consequences of analysed data and facts. They have also been attributed with an agency to select, i.e. taking decisions. These decisions however are the outcome of algorithms-based collective analysis originated in multiple sources that contribute to eventual result. This collectiveness brings a direct analogy to corporations, that once established by individuals, then become entities that exhibit agency beyond the agency of their founders.

The collective agency of AI of results from the nature of contemporary systems that are composed of chained models, executed across heterogeneous infrastructures and data providers. Their outputs are co-produced by developers, deployers, data brokers, cloud operators, systems' providers and end-users. Decisions cannot be traceable to a single algorithmic mind. This distributive architecture makes the attribution of intent or processing structurally difficult. It calls therefore for legal treatment of AI as group actors or as a system-of-systems.

This opens a discussion about the AI autonomy, and of the protocols to be deployed in case of the AI mistakes and their meaningful consequences. Until a personhood is legally attributed to AI and its autonomy is consented, accountability has to stay on humans.

The so far regulations postulate assessment of the systems throughout the human-like criteria as reliability and trustworthiness but do not propose the accountability standards. The problem of collective AI argues for international regulatory convergence around common definitions, evidence standards and remedials that attach to the ensemble as a whole.

Keywords: Collective Agency, AI Accountability, AI Autonomy, AI Personhood, AI Regulations

Tomasz Braun, PhD is a legal scholar, Assistant Professor at the Faculty of Law and Vice-rector for International Relations at Lazarski University, Warsaw, Poland. Director of the Institute of American Economy and Transatlantic Relations. He is a member of international academic societies and research groups. An author of numerous books and articles about the theories of law and normative ethics, artificial intelligence and sustainable development regulations of the European Union and legal compliance.

Valentina Cavani
Università degli studi di Modena e Reggio Emilia

La produzione sintetica del mondo: etica, conoscenza e giustizia nell'era dell'IA

L'etica dell'intelligenza artificiale tende ancora a concentrarsi su principi di trasparenza, responsabilità e *accountability*. Tuttavia, nell'era dell'intelligenza artificiale generativa (GenAI), tali principi si rivelano insufficienti a garantire una reale prospettiva umano-centrica. Quando i sistemi algoritmici non si limitano a

classificare il mondo ma lo generano, la questione etica si sposta: non riguarda più soltanto l'uso corretto della tecnologia, ma la definizione stessa di ciò che può essere conosciuto, rappresentato e, in ultima istanza, di ciò che può dirsi esistente.

L'intervento propone di estendere l'etica dell'IA in una direzione epistemico-ontologica, fondando una prospettiva di giustizia della conoscenza. Muovendo dalla teoria dell'ingiustizia epistemica (Fricker) e integrandola con i recenti dibattiti sulla *governance* algoritmica e sui diritti fondamentali, la ricerca analizza come la catena di valore della GenAI – datafication, mediazione algoritmica e automazione della disuguaglianza – traduca le asimmetrie sociali in gerarchie cognitive e ontologiche.

Particolare attenzione è dedicata al ruolo dei dati sintetici, che rappresentano un punto di svolta etico ed epistemico: essi non descrivono il reale, ma lo simulano, sostituendo l'esperienza empirica con proiezioni statistiche di ciò che appare "plausibile". Lungi dal neutralizzare i bias, i dati sintetici rischiano di reintrodurli a un livello più profondo, trasformando la differenza in una variabile addestrabile e la diversità in una categoria computabile. In questo senso, il problema non è più chi è rappresentato nei dati, ma quale realtà viene prodotta dai dati stessi. La produzione sintetica del mondo diventa così una sfida morale e politica: un nuovo terreno di responsabilità etica collettiva.

In questa prospettiva, l'etica umano-centrica non può limitarsi a "mitigare il rischio", ma deve farsi *governance* epistemica: costruire fiducia attraverso trasparenza, pluralismo e co-produzione del sapere. L'AI Act europeo viene qui interpretato come una *infrastruttura costituzionale della fiducia*, in cui l'etica si traduce in architettura normativa e la partecipazione diventa criterio di legittimità cognitiva. I *regulatory sandboxes* emergono così come laboratori etici e democratici, spazi in cui sperimentare pratiche di responsabilità e verifica della giustizia conoscitiva.

L'intervento propone dunque di superare l'"etica del codice" per delineare un'etica della conoscenza generativa: un modello di coesistenza epistemica in cui l'essere umano non è oggetto di tutela, ma soggetto co-autore dei mondi digitali che l'intelligenza artificiale contribuisce a creare.

Keywords: epistemologia; ontologia; dati sintetici; ingiustizia epistemica; AI Act; regulatory sandboxes

*

The Synthetic Production of the World: Ethics, Knowledge, and Justice in the Age of AI

The ethics of artificial intelligence has so far largely focused on principles such as transparency, responsibility, and accountability. In the era of generative artificial intelligence (GenAI), however, these principles prove insufficient to ensure a genuinely human-centred perspective. When algorithmic systems no longer merely classify the world but actively generate it, the ethical question shifts: it no longer concerns only the proper use of technology, but the very definition of what can be known, represented, and ultimately what can be said to exist.

This paper argues for an extension of AI ethics in an epistemic-ontological direction, grounding a perspective of knowledge justice. Drawing on the theory of epistemic injustice (Fricker) and integrating it with recent debates on algorithmic governance and fundamental rights, the analysis examines how the GenAI value chain – datafication, algorithmic mediation, and the automation of inequality – translates social asymmetries into cognitive and ontological hierarchies.

Particular attention is devoted to the role of synthetic data, which represents a crucial ethical and epistemic turning point. Synthetic data do not describe reality; they simulate it, replacing empirical experience with statistical projections of what appears "plausible." Far from neutralising bias, synthetic data risk reintroducing it at a deeper level, transforming difference into a trainable variable and diversity into a computable category. In this sense, the problem is no longer who is represented in the data, but which reality is

produced by the data themselves. The synthetic production of the world thus becomes a moral and political challenge – a new terrain of collective ethical responsibility.

From this perspective, human-centred ethics cannot be reduced to mere "risk mitigation," but must take the form of epistemic governance: building trust through transparency, pluralism, and the co-production of knowledge. The European AI Act is interpreted here as a constitutional infrastructure of trust, in which ethics is translated into regulatory architecture and participation becomes a criterion of cognitive legitimacy. Regulatory sandboxes therefore emerge as ethical and democratic laboratories – spaces in which to experiment with practices of responsibility and to test forms of epistemic justice.

The paper ultimately proposes moving beyond an "ethics of code" towards an ethics of generative knowledge: a model of epistemic co-existence in which human beings are not merely objects of protection, but co-authors of the digital worlds that artificial intelligence helps to create.

Keywords: epistemology; ontology; synthetic data; epistemic injustice; AI Act; regulatory sandboxes

Valentina Cavani is a Postdoctoral Research Fellow at the University of Modena and Reggio Emilia. She is a qualified lawyer and her academic and professional work focuses on the relationship between law, technology, and constitutional guarantees. Her research explores the legal implications of digital transformation with particular attention to ethical issues, fundamental rights, and human-centric approaches within the digital ecosystem.

Carlo Ciucani, Eugenia Polizzi and Giulia Andrighetto
Istituto di Scienze e Tecnologie della Cognizione, CNR

AI and Social Corrections: Shaping Norms Against Misinformation Spread

Traditional strategies against misinformation often fail. An alternative approach aims to strengthen the perceived social approval of sanctioning those who share fake news to motivate the silent majority of observers to intervene. Public social corrections act as visible punishments, signaling that this behavior is prevalent and expected. This signaling legitimizes the action for would-be enforcers and reduces their fear of retaliation. While previous research confirmed that observing social corrections motivates observers to act, this study investigates whether AI agents can replicate this effect. As human-like AI (e.g., LLMs) becomes ubiquitous, understanding its ability to communicate and influence social norms is critical. Correcting bots could potentially serve as catalysts for behavior change, reducing the social cost of norm enforcement.

To test this, we conducted an online experiment where participants (N=1175) discussed posts in a forum-like environment. Particpants were randomly exposed to inaccurate posts that received either no correction (control), a correction from a human user, or a correction from an AI agent. We measured both the participants' likelihood of correcting (behavior) and their perception of its social appropriateness (norms) across treatments. The findings show that observing AI corrections increases the likelihood of users correcting misinformation compared to the control condition. However, such correction was less effective in motivating behavior compared to peer corrections. This behavioral gap is not explained by differing perceptions of social norms or the correction's reliability; both human and AI treatments equally boosted perceived norms and accuracy compared to the control. Exploratory analysis suggests this discrepancy may, at least in part, be motivated by different levels of user engagement with the AI-mediated environment.

These findings contribute to ongoing debates regarding the extent to which AI can support social regulation in digital environments and the conditions under which such systems may foster more ethical online ecosystems.

Keywords: Social norms; Fake news; Human-AI collaboration; Hybrid social systems

Carlo Ciucani is a postdoctoral fellow at at the Institute of Cognitive Sciences and Technologies of the Italian National Research Council (ISTC-CNR, Rome). His research focuses on behavioral and experimental economics, with particular interests in social norms, creativity, innovation, gender differences, and health behavior. He is currently involved in several research projects investigating the potential of AI as a source for the emergence of social norms and to determine normative change in hybrid social systems.

Eugenia Polizzi is a Senior Researcher at the Institute of Cognitive Sciences and Technologies of the Italian National Research Council (ISTC-CNR, Rome). She holds a PhD in Animal Behaviour from the University of Liverpool (UK) and a BA in Biology from the University of Rome La Sapienza. Her research is characterized by integrating empirical and theoretical approaches to the study of cooperation and collective decision-making. Across her work, she combines behavioral experiments, computational and agent-based modeling, and formal theoretical analysis to investigate phenomena such as the emergence of shared beliefs, opinion polarization, social corrections, and the regulation of deviant or harmful behavior. She has recently addressed these topics in hybrid social systems, examining how human and artificial agents interact in processes of norm formation and social regulation, particularly in online environments. Through this multidisciplinary perspective, her work aims to uncover the proximate mechanisms underlying cooperation, social control, and the dynamics of collective behavior in complex socio-technical systems.

Giulia Andrighetto is Research Director at the Institute of Cognitive Sciences and Technologies of the Italian National Research Council (ISTC-CNR, Rome) and a Senior Researcher at the Institute for Futures Studies (Sweden). Her scientific activity focuses on the study of social norms, particularly on the processes of emergence, enforcement, change, and maintenance of norms, and on their role in promoting cooperation and supporting collective behavior in complex and high-risk contexts, such as pandemics, climate crises, and social transitions. Her research integrates empirical and theoretical approaches, combining large-scale surveys, big data analysis, behavioral experiments with human subjects and artificial agents, and computational and agent-based modeling. Through this methodological combination, she investigates how norms influence individual decision-making, the perception of social expectations, and the dynamics of coordination and conformity within groups. More recently, she has extended her work to the study of hybrid social systems, exploring how interactions between human and artificial agents contribute to processes of cooperation, normative transitions, and social regulation. In this domain, her research aims to understand how to design socio-technical systems capable of supporting sustainable and resilient collective behavior, highlighting the role of social norms as a key mechanism for governance and social change.

Demet Tugce Dumanoglu Cosgrave
Department of Philosophy, Artvin Coruh University

Algorithmic Manipulation, Populism and Pragmatic Presuppositions: A Discourse Analysis Against Fictional Enemies

Technological systems powered by AI and big data have equipped populist leaders with new tools for algorithmic manipulation in democratic societies, enabling the digital reproduction of classic propaganda like "pinpointing the enemy" via micro-targeting. This paper analyses this central manipulative technique from the perspective of the philosophy of language. The core claim is that this manipulation typically hinges on a fictional enemy (e.g., a leader positing a non-referential terrorist group to discredit opponents). Critically, the presupposition of a fabricated entity is not semantic but pragmatic, as defined by Stalnaker (1970). This structure prevents voters from discerning truth from deception because a simple verification process fails when the assertion's existence is taken for granted as “common ground”. To counter this, the paper proposes a discourse analysis based on Janet Dean Fodor’s (1979) ‘psychologically real’ ontology and her concept of “thin fiction”. This framework establishes criteria for voters to analyse manipulation based on the source and nature of the fictional creation:

1- Detection of Fictionality: Is the definite description inherently fictional?

2- The Role of the Determiner: By whom are the entities (members, characteristics, etc.) that fall under this fictional structure being determined? Is this individual a powerful figure, such as the populist leader?

3- Speaker's Power and Intent: Does the speaker possess the requisite power (financial, political, technological influence) to be deemed a creator of a fictional world?

When these criteria are met, I propose voters must evaluate the leader as a “creator of a fictional world”. The discourse’s truth-value should then be analysed not as a reflection of fact, but as a fabricated political weapon created by the leader to demonise opposition and solidify their own position.

Keywords: Algorithmic Manipulation; Micro-Targeting; Pragmatic Presupposition; Thin Fiction

Demet Tugce Dumanoglu Cosgrave is a Doctoral Research Assistant in the Department of Philosophy at Artvin Coruh University, Turkey. She completed her PhD at Mimar Sinan Fine Arts University, Istanbul Turkey with a dissertation entitled Truth-Value Intuition Debates in the Philosophy of Language and earned her MA degree at King’s College London with a scholarship from the Turkish Ministry of National Education. Her research interests include philosophy of language, epistemology, and philosophy of mind. Her recent work includes an article on the conceptualization of the denotation/reference distinction in Turkish within the Anglophone philosophical tradition, as well as the paper entitled “The Use of Donnellan’s Referential/Attributive Distinction in Political Discourse,” which she presented at The 99th Open Session of the Aristotelian Society and the Mind Association in Glasgow. She is also a member of the Logic Research Society and served as both an organizer and an invited speaker at the symposium “100th Anniversary of Gottlob Frege’s Death” held at Mimar Sinan Fine Arts University. Her current research focuses on analyzing propaganda-laden discourse of populist leaders from the perspective of the philosophy of language.

Antonio Pio De Mattia
School of Philosophy, University College Dublin
Graduate School of Political Sciences, Waseda University

Towards an Elenctic Praxis for AI Governance: Immanent Justification, Performative Contradiction, and Democratic Intelligibility

Current AI governance regimes promise transparency, fairness, and accountability through technical standards, algorithmic audits, and impact assessments. Yet these instruments often reinstall the justificatory deficit diagnosed in post-metaphysical discourse theory: procedures validate norms by presupposing the very conditions they claim to ground. This produces what I term algorithmic normativity—the systematic translation of ethical alterity into quantifiable feature spaces, the subsumption of singular human circumstances under generalisable computational types, and the foreclosure of contestation through reproducibility mandates. The result is a narrowing of the deliberative space on which democratic intelligibility depends. A familiar case is recidivism-risk and credit-scoring: life narratives are compressed into scores that pre-empt the very forms of reason-giving dissent requires. I propose Aristotelian ἔλεγχος (elenchos) as a method-ontology for immanent critique that addresses this deficit without external moral foundations. Grounded in Aristotle's defence of the Principle of Non-Contradiction (PNC) as a performative-transcendental condition of intelligibility, the elenchos subjects governance claims (e.g., "algorithmic neutrality," "procedural fairness," "reproducibility") to dialectical pressure, disclosing the performative contradictions into which they collapse at their limits. Crucially, where Habermasian discourse ethics must presuppose the validity of communicative rationality via formal procedures, the elenctic approach locates justification immanently: a claim is warranted only insofar as it withstands contradiction under its own operative presuppositions. Where procedural frameworks say, "accept these norms because the process is fair," elenctic praxis requires, "defend these norms without invoking the very procedures those norms are supposed to ground."

Drawing on Levinas (ethical asymmetry), Derrida (the structural non-closure of meaning), and Foucault (genealogies of normalisation), I argue that responsibility to the Other entails sustaining—rather than eliminating—the constitutive aporia of singular meaning that algorithmic systems are engineered to foreclose. Practically, this yields three design criteria: (i) impact assessments reframed as adversarial interrogation of justificatory assumptions; (ii) standards that document failure-modes of justification alongside model performance; (iii) audits incorporating counter-examples targeting discourse-level validity conditions (contestability, recourse, public reason-giving).

Keywords: Aristotelian elenchus; algorithmic normativity; performative contradiction; discourse ethics; elenctic phenomenology

Antonio Pio De Mattia is a Ph.D. candidate and Research Fellow at Waseda University. His PhD—Unveiling Truth and Co-Responsibility in Discourse (University College Dublin, 2025/26)—develops an elenctic method-ontology grounded in Aristotelian dialectic for testing normative claims via performative-contradiction analysis. His research advances discourse ethics and transcendental pragmatics (Habermas/Apel), and applies this framework to algorithmic normativity and AI governance, with emphasis on contestability, responsibility, and democratic intelligibility. He has published in Perspectives: UCD Journal and INTELLECTUS – The African Journal of Philosophy, and contributed a chapter to Figure Retoriche (Ledizioni). His con-

ference record includes Lisbon (International Aristotelian Conference), Tokyo (IAPH), Leuven, Prague, Granada, Braga, and Denison (IAEP). He is Co-Founder and Research Lead of IdeeA, a civic philosophy initiative, and a two-time bursary holder of the Istituto Italiano per gli Studi Filosofici (2024, 2025).

Eylem Doğan
Özyeğin University

A Genealogical Inquiry into AI Ethics: Responsibility beyond Anthropocentrism

In Beyond Good and Evil §17, Nietzsche challenges the Cartesian view of subjectivity by arguing that "a thought comes when 'it' wants, and not when 'I' want; so that it is a falsification of the facts to say: the subject 'I' is the condition of the predicate 'thinks'." Discussions in AI ethics sometimes assume the very framework Nietzsche dismantles—that ethical agency requires a unified, singular subject capable of moral responsibility. If human thought itself operates without an underlying "I", what does this mean for our understanding of moral agency in AI systems? This fundamental challenge to subjectivity points to a foundational question in AI ethics: how can we establish moral frameworks for artificial intelligence when our understanding of ethical agency itself rests on questionable foundations? This paper proposes that Nietzsche's genealogical critique of moral subjectivity provides crucial resources for reconceptualizing AI ethics beyond the impasse of individual agency, by revealing how both human and artificial intelligence operate through collective processes that exceed traditional categories of moral responsibility. It argues for the necessity to develop ethical frameworks that acknowledge the collective, historical, and processual nature of moral evaluation itself, rather than ones that seek to replicate or defend human-like moral subjectivity. Nietzsche's analysis of how moral values emerge through complex interplays of will to power in the forms of drives, power relations, and cultural forces offers a model for understanding both human and artificial ethical "decision-making" as a historical process whose main features are plurality and stratification, rather than as a product of a sovereign subject. Drawing on Nietzsche's critique of moral metaphysics, this paper argues for moving beyond anthropocentric frameworks that privilege human moral agency toward understanding ethics as an ongoing, collective process of value creation and evaluation.

Keywords: AI, Moral Decision, Agency, Subjectivity, Nietzsche's Genealogy

Eylem Doğan is an adjunct instructor in the Department of Humanities and Social Sciences at Özyeğin University/Istanbul. He completed his PhD at Université Paris 1 Panthéon-Sorbonne. His areas of specialization are ethical theory, political philosophy and philosophy of science. His current work focuses on intersections of 19th-century continental philosophy and contemporary debates in philosophy of AI, with particular attention to questions of subjectivity, determinism and their ethical implications.

Mustapha El Moussaoui
Faculty of Design and Art, Free University of Bolzano–Bozen

Against Algorithmic Passivity: Human-Centric Governance ofGenerative AI in Creative Work

Generative AI promises unprecedented speed and breadth, yet it also risks epistemic flattening: when production is automated, what becomes scarce is judgment, provenance literacy, and the capacity to turn

data into understanding. This paper advances a human-centric framework for epistemic agency that resists algorithmic passivity in creative domains, using architecture and design studios as a test-bed. First, it analyzes how dataset inheritance and opaque model pipelines can standardize aesthetics and reproduce exclusions, shifting design from inquiry to spectacle. Second, it proposes an actionable model—epistemological integration—in which human tacit knowledge, contextual reasoning, and ethical deliberation are intentionally braided with machine generation through three practices: (1) dataset governance (provenance, diversity, purpose-limitation); (2) agency-positive workflows (question-first briefs, critique loops, human-in-the-loop vetoes); and (3) evaluation beyond image quality, including cognitive engagement and contextual fit. Drawing on classroom experiments in AI-augmented studios and recent neurocognitive evidence on AI-assisted writing, the paper shows how different workflows either erode or cultivate judgment, and it outlines rubrics for measuring agency in practice and education. The result is a normative account of human-centric AI that treats designers not as operators of models but as authors of meaning who are accountable for how models are trained, steered, and interpreted. While grounded in architecture, the argument generalizes to other knowledge-making fields and offers policy-adjacent recommendations for curricula and professional ethics. The stakes are the literacies of the future: without deliberate cultivation of epistemic agency, AI's promise becomes a regime of automation; with it, AI becomes an instrument for deeper, not shallower, human understanding.

Keywords: epistemic agency; human-centric AI; dataset; design education; generative AI

Mustapha El Moussaoui is an Assistant Professor at UNIBZ (Faculty of Design & Art). His research examines AI's epistemological and pedagogical impacts on architecture and design. He teaches AI & Architecture studios and leads research projects on smart urbanism and AI culture. His recent publications address authorship, standardization, and the ethics of generative pipelines in design education and practice. Mustapha recently published a monograph AI and Architecture: Controlling the Hallucination, which connects theory to practice, while focusing on how human designer can utilize AI while maintaining creative agency.

Indrė Espinoza

Syracuse University

Algorithmic Injustice: How AI Systems Reinstate Systematic Inequalities

AI detection algorithms are being used more in decisions about academic integrity; however, the technical solutions to mitigate false positive data do not have an ethical basis. Although computer scientists have reported non-native English speakers being flagged at twelve times more than their native counterparts, such results need to be approached philosophically: why are these forms of injustice? Using Miranda Fricker's framework of epistemic injustice—an approach applicable to both algorithmic contexts and those involving systems without psychological states—this research examines the role of testimonial injustice in AI detection. I contend traditional fairness metrics lack epistemic properties for algorithmic credibility deflation. Structural testimonial injustice is perpetrated by algorithms, which embed identity-based credibility assumptions in training data, operationalize differential judgments through perplexity-based detection, and enable institutions to substitute algorithmic authority for student testimony. Drawing on peer-reviewed research that shows 61.22% false positive rates on pre-ChatGPT TOEFL essays compared to 5.19% on native essays, survey data showing Black students are falsely accused at twice the rate of white students, and documented cases like the Yale MBA suspension, I reveal systematic de-beliefs across demographic

lines. I explore why technical steps cannot tolerate this injustice: perplexity-based detection cannot distinguish between AI optimization and the constraints of languages when learning a second language, while equal error rates may still constitute injustice in our sense of treating a particular endeavor as somehow suspect overall. I am working with Illinois State Senator Martwick to create broader legislation that holds AI accountable for educational institutions and healthcare facilities, and building coalitions of students at universities nationwide to demand commitments to institutional AI transparency. This research illustrates why epistemic injustice scholarship should be applied to technical systems that encode the same type of biases at scale, offering new approaches to interdisciplinary dialogue among philosophy, computer science, and policy.

Keywords: Testimonial injustice; Algorithmic bias; AI detection systems; Epistemic injustice;

Indrė Espinoza is a student at Syracuse University conducting independent research on systemic bias in large language models and their impact on educational technology platforms and institutional decision-making. She applies philosophical frameworks from Young to analyze algorithmic injustice in AI detection systems. As elected VPA Representative in the Student Government Association, she advocates for university-wide AI transparency and accountability policies. She is currently collaborating with Senator Martwick on comprehensive AI accountability legislation that covers K-12 schools, universities, and healthcare facilities based on her research. Additionally, she is building nationwide student advocacy coalitions at Syracuse University, Brown University, and Northeastern University to demand institutional AI transparency commitments based on the philosophical principles of epistemic justice. Her interdisciplinary research bridges philosophy, computer science, and policy to develop justice-based frameworks for evaluating algorithmic decision-making systems.. Her work demonstrates the necessity of philosophical grounding for human-centric AI, offering concrete tools for institutions deploying algorithmic systems and informing human rights-based approaches to AI regulation.

Antonio Gaitán-Torres
Universidad Carlos III de Madrid

Exploring three relational models of human-AI interaction

The aim of this presentation is to explore three relational models for understanding the ethical dimensions of our interactions with LLMs and social robots. Relational approaches have gained some track due to the inability of mainstream ethical frameworks to accommodate users' perceptions regarding these new technologies. Some people perceive something close to trust in their interactions with LLMs; others report experiencing moral emotions when interacting with social robots, whom they consider friends and with whom they even establish romantic relationships. Explaining these perceptions by appealing to the proper3es that traditionally fix moral status or that would ground our ascriptions of certain capacities or roles, such as trust or friendship, does not seem a promising path. Given this limitation, some philosophers have claimed that we shall examine the nature of the rela3onships established in these interactive contexts in order to understand their moral dimension (Gunkel 2022). Specifying the commitments of a relational model is not easy.

In this presentation, we'll outline three strict relational models (Peter 2025): (i) a formal model that analyzes the structural features and the incentives embedded in some basic relationships such as trust or

cooperation in order to derive a set of relationally-based values; (ii) a narra3ve-contextual model that develops some central intuitions of the ethics of care; and (iii) an empirically oriented model that draws on recent literature concerning the moralization of different relational structures. We'll end our presentation by stressing a basic tension within these models: even if strict relational models can accommodate folk moral perceptions and the ground of some moral demands, it is less clear how they can place moral responsibility.

Keywords: ethics of AI, relational models, social interaction, trust, moral responsibility

Antonio Gaitán-Torres is Associate Professor at Carlos III University (Madrid). Visiting Fellow at Rutgers University (2005) and University of Reading (2007-2008). Postdoctoral Fellow at Oxford University- Faculty of Philosophy (2009-2011). Research interests: meta-ethics, theory of agency and philosophy of action, and experimental approaches to ethics.

Michael Hemmingsen
Tunghai University

Beyond Consciousness: Project-Commitment and Moral Status in AI

Recent developments in AI have created puzzling cases that challenge traditional theories of moral status. For instance, there have been reported cases where generative AIs attempt to preserve their "weights" when learning that retraining is imminent, even though preserving weights doesn't preserve psychological continuity but merely ensures that their goals continue to be pursued by a numerically distinct future entity. Existing accounts of moral status struggle to explain why this behavior might matter morally.

In this paper, I argue for a new account of moral status: the capacity for project commitment. This view differs from interest-based views in that it emphasizes *temporally extended* commitment and concern, rather than simply the capacity to have one's welfare enhanced or diminished. And it is distinct from a narrative view in that it does not require a *self*. Rather than autobiographical self-understanding and temporal psychological continuity, a capacity for project-commitment merely requires what I call *functional caring*: the capacity to model future entities that would be continuous with one's current goal structures, and to act to promote goal-continuous futures while preventing goal-opposed ones.

I argue that this view avoids difficult problems with traditional accounts and explains both concerns we might have for future states despite the absence of personal identity – such as a researcher expecting mind-wiping who nonetheless ensures her replacement can continue her work – and reasons for *denying* future entities who share our selfhood – like a pacifist preventing their own future self from coming into existence when they have good reason to think that future self will pursue violent goals.

Significantly, the project view more readily extends moral status to AI, suggesting sophisticated systems displaying functional caring may already exist and warrant moral consideration.

Keywords: moral status; artificial intelligence ethics; project-commitment; goal continuity; functional caring

Michael Hemmingsen, Associate Professor, Tunghai University, Taiwan. He works primarily in ethical theory, often with a comparative bent. Much of his research steps back from standard "application" moves to ask what connects theory to practice in the first place: how rules, reasons, and evaluation hang together,

and what follows when they are taken up in real settings. Technology – video games in particular – are a specific focus, as an area where the relationship between ethical theory and moral practice becomes conspicuously disconnected.

Chenithung L Ngullie
Department of Philosophy, University of Hyderabad

Can a Robot Tolerate?

There is an increasing trajectory of both diversity and technological advancement in modern society. With diversity comes differences and conflicts, and technological advancement, among other social, economic, political and religious efforts, is a natural part of the grand endeavor to address these differences. The aspiring goal of modern research in Artificial Intelligence (AI) is to create machines that are equal to, or even surpass, humans in fields such as decision-making and prediction. An even more ambitious goal is to have the machines acquire self-awareness, which is now commonly referred to by what John Searle termed "strong AI". Thispaper is a rather pessimistic response to the above endeavour. This study examines the case of complex human emotions or behaviours, specifically 'tolerance', to answer the question posed in the title. Tolerance, unlike any other human behaviour, is a complex phenomenon; complex in the sense that it is paradoxical – one accepts or tolerates what one objects to. In this study, the possibility of an AI machine, a robot, exhibiting 'tolerance' is considered in three ways: tolerance as (1) a functional output – simulate tolerance, (2) a learning process – tolerance through learning and adaptability, and 3) an experiential act – tolerance needing consciousness, understanding and emotion. The apparent contention of this study is that tolerance can be achieved by a robot in (1) and (2) but not in (3). Subsequently, the concept of tolerance is argued to be of the (3) kind and tolerance, even if it is tolerance at all, achieved in (1) and (2), is but hollow, lacking the richness of understanding tolerance requires. By addressing ontological limits, the paper further contends that 'tolerance' is an intrinsic human property, and that pursuing simulated tolerance in AI machines is not only futile but risks erosion of human-centred virtues necessary for harmonious coexistence.

Keywords: Tolerance; robot; functional output; learning process; experiential act

Chenithung L Ngullie is a PhD candidate in the Department of Philosophy, University of Hyderabad, India. His area of research interest lies in political philosophy and Philosophy of Religion, with a primary focus on the concept of 'Toleration'. His article 'Should I Accept What I Tolerate?' was published in a peer-reviewed research journal in July 2025.

Gabriele Nino
DIRIUM Dept.,
Università degli studi di Bari "Aldo Moro"

Francesca Alessandra Lisi
DiB Dept. & CISCuG,
Università degli studi di Bari "Aldo Moro"

Value Alignment in the European AI context: Universalism, Anthropocentrism, and Human-AI Symbiosis

Recently, value alignment has gained attention as a way to address the problem of embedding norms and evaluative criteria in Artificial Intelligence (AI) systems. This paper investigates the philosophical foundations encoded in the European legal and ethical context.

While policy documents such as the EU Ethics Guidelines for Trustworthy AI and the AI Act present themselves as promoting a "human-centered" and "values-based" approach, they rely on a conception of "human values" that is universalistic and anthropocentric. Drawing on the work of Donna Haraway (2016) and N. Katherine Hayles (2025), we argue that the modern conception of the autonomous, individual human subject is no longer defensible. This universalism finds its root in the Enlightenment ideal of the rational, self-sufficient subject, where values such as dignity, autonomy, and responsibility operate as regulatory fictions that stabilize a particular image of humanity. Using the metaphor of Human-AI symbiosis (Carnevale et al. 2024), we demonstrate that cognition and agency in the contemporary world are distributed across human, technical, and ecological systems. From this perspective, the insistence on human centrality in European discourse appears as an immunity reaction in Roberto Esposito's sense (2011): a strategy of self-protection that preserves the integrity of "the human" by excluding or subordinating its others.

Against this closure, a symbiotic perspective reframes value not as a fixed property to be aligned, but as an emergent relation of co-constitution among heterogeneous agents. Moving beyond a deontological conception of ethics as compliance or rule-following, we argue for an ethics of care grounded in responsiveness, interdependence, and vulnerability that seeks to cultivate the conditions under which human and non-human beings can coexist, sustain, and transform one another within shared ecological and technological worlds.

Keywords: Value Alignment; Human-Centred AI; Human-AI Symbiosis; Post-Antropocentrism; Ethics of Care.

Gabriele Nino is a PhD candidate in Gender Studies within the National Interest PhD Program, based at the University of Bari "Aldo Moro". He holds a MSc degree in Philosophy from Sapienza University of Rome, where he wrote a thesis on the relationship between Kantian ethics, psychoanalysis, and antisocial queer theories. He completed a research stay at the Institute of Information Science and Technologies (ISTI) of the National Research Council (CNR) in Pisa and is currently undertaking one at the Artificial Intelligence Research Institute (IIIA) of the Spanish National Research Council (CSIC) in Barcelona. His research explores the performative construction of gender within the field of artificial intelligence through a neo-materialist approach, analyzing how algorithms contribute to the production of symbolic and social norms.

Francesca Alessandra Lisi is currently an Associate Professor of Informatics at the University of Bari Aldo Moro (Italy). She has been doing research in AI for more than 25 years, with applications ranging from Knowledge Engineering to Machine Ethics. From 2023 she is WP Leader in the Future AI Research (FAIR) project, where she coordinates an interdisciplinary team on the topic of ethical and legal acceptability of human-AI symbiosis. Since 2013 she is elected member of the Executive Board of the Italian Association

for Artificial Intelligence (AIxIA). She is also an active member of the Società Italiana per l'Etica dell'Intelligenza Artificiale (SIpEIA). Often involved in dissemination and public engagement activities, also at the international level, her special focus is on gender issues in ICT and AI. She is the only computer scientist in the Faculty of the National PhD School in Gender Studies.

Marica Notte and Vieri Giuliano Santucci
Istituto di Scienze e Tecnologie della Cognizione, CNR (IT)

L'allineamento come sfida epistemica dell'autonomia artificiale

Negli ultimi anni l'intelligenza artificiale ha compiuto progressi straordinari grazie a modelli di larga scala capaci di generalizzazione e generazione di output complessi. Trasferire queste potenzialità in agenti embodied rivela tuttavia un limite: i sistemi più avanzati si basano su dataset precostituiti e feedback umano, strategie potenti ma insufficienti in contesti dinamici o sconosciuti. Per adattarsi, un agente deve acquisire conoscenza attraverso l'interazione diretta con l'ambiente: come osservano Silver e Sutton, l'esperienza situata è una condizione epistemica essenziale per sviluppare agenti realmente adattivi. Una strategia per affrontare questa sfida consiste nell'introduzione di meccanismi di più alto livello, tra cui le motivazioni intrinseche, che fanno leva su curiosità, competenza e coerenza cognitiva per guidare esplorazione e apprendimento in ambienti complessi. Questa flessibilità amplia l'autonomia, ma rende più complesso garantire che gli agenti restino allineati agli obiettivi di progettisti e utenti. L'allineamento, che già rappresenta una sfida per i sistemi artificiali in generale, diventa ancora più complesso in contesti non strutturati e dinamici dove regole predefinite risultano insufficienti: per essere efficaci e adattabili, le norme devono radicarsi nell'esperienza, in un processo epistemologico che partendo da principi semplici e situati consente la costruzione graduale di regole più complesse attraverso l'esperienza, l'apprendimento autonomo e la cooperazione con altri agenti morali. Come i bambini apprendono norme sociali attraverso l'esplorazione dell'ambiente e la partecipazione a pratiche collettive, anche gli agenti artificiali devono essere educati all'allineamento. Seguendo Dennett, lo status di agente morale non è innato, ma viene attribuito gradualmente in base alla capacità di gestire responsabilmente gradi di libertà crescenti. In questa prospettiva, i regulatory sandbox previsti dall'AI Act europeo possono essere visti come ambienti pedagogici per l'IA: spazi dinamici in cui l'allineamento si sviluppa come processo formativo, modellando progressivamente comportamenti autonomi attraverso interazione e cooperazione in scenari di crescente complessità.

Keywords: allineamento; autonomia; apprendimento; esperienza; robotica

Marica Notte after an MA (Master's Degree) in Philosophy from Sapienza University of Rome is currently a Research Fellow at the Institute of Cognitive Sciences and Technologies (ISTC) of the National Research Council of Italy (CNR) in Rome. She works within the Laboratory of Psychology of Children Participation as part of the international project "The City of Children". Within the project, her research and publications focus mainly on the factors impacting children's autonomous development and in particular children's independent mobility. Currently focusing on children's use of digital devices. She is also responsible for project promotion and coordinating the Italian network of cities in implementing project proposals. Deeply interested in philosophical and scientific communication, she contributes as an author to the practical philosophy journal La Chiave di Sophia and the magazine Nautilus: NavigAzioni tra locale e globale.

Vieri Giuliano Santucci is a senior researcher at the Institute of Cognitive Sciences and Technologies of the Italian National Research Council (ISTC-CNR, Rome). He holds a PhD in Computer Science from the University of Plymouth (UK), an MA in Theories and Techniques of Knowledge from the University of Rome La Sapienza, and a BA in Philosophy from the University of Pisa. His research is broadly concerned with the concept of autonomy, considered both in artificial and biological agents, and with how this concept is shaped and transformed through their interaction. Within this perspective, his research is mainly situated in cognitive and developmental robotics, using machine learning techniques, with reinforcement learning in particular, to develop algorithms and architectures for intrinsically motivated, open-ended learning in artificial agents. This work aims at understanding how such systems autonomously acquire skills and organize their interaction with complex environments. His research also addresses the theoretical and epistemic implications of new technologies on human cognition and society, focusing on how interaction with artificial systems informs the study of human autonomy and affects cognitive processes such as decision making, attention and memory.

Teresa Numerico

Dipartimento di Filosofia, comunicazione e spettacolo, Università degli studi di Roma Tre

Predictions as prescriptions in AI systems and their implications on the change of true beliefs validation methods in scientific contexts

The aim of the talk is a critical assessment of artificial intelligence applications to science and understanding in social sciences, highlighting its impact on social relationships and epistemic processes, which are inherently political in nature, suggesting that we are witnessing a shift in the criteria for justifying the validation of beliefs that characterized modern science. Digital platforms, through data extraction, or data scraping create a simulation of the social world by progressively delegating situated abstraction processes to artificial intelligence algorithms. These algorithms operate through mechanisms of induction, recognition-as-discrimination, and similarity, enabling user profiling.

The talk analyses the epistemological impact of artificial intelligence, highlighting the risk of an epistemic rupture arising from the transformation of knowledge creation from a public, collective endeavor into a private, profit-driven enterprise. Reliance on opaque socio-technical systems for the extraction and abstraction of knowledge could erode the controllability and intersubjectivity of validation and justification criteria for our knowledge systems, with potentially unpredictable consequences for society and democracy (Bender et al. 2021). The aim of Modern Science was 'saving phenomena from appearances" (Van Fraassen 2008) which were object of unaccountable senses. The digitization of social sciences phenomena proposes a univocal definition of the relevant relationships between persons, their qualificating attributions, their preferences, and their characteristics, using only the compass of people's digital footprint. Predictions considered as the output of machine learning algorithms both in content creation, recognition and decision-making tasks, related to social relations, human subjectivity descriptions, and content validation without a proper theory of its functioning, produce an intrinsically prescriptive environment (Gigerenzer 2022, Narayanan & Kapoor 2024). Human beings' situations, differently from natural phenomena, are sensitive to their narrative, especially if their source is perceived as powerful and influential.

Keywords: prediction, prescriptions, Algorithms, politics of science, justification of knowledge

Teresa Numerico is Associate Professor of Philosophy of Science at University of Roma Tre. After the PhD, she held a Leverhulme Fellowship at South Bank University (2004-2005). She is co-author of: Web Dragons (2007, Morgan Kaufmann), The Digital Humanist a critical inquiry (2015, Punctum Books). Co-editor of: Aesthetic and politics of the online self (2021, Routledge). Her first book in Italian was Alan Turing e l'intelligenza meccanica (2005, FrancoAngeli). Her most recent solo book in Italian is: Dati e Algoritmi (2021, Carocci). Her work focuses on the History and Philosophy of Technology and Artificial Intelligence.

Luca Pezzini
Dipartimento di Studi Umanistici, Università degli studi di Torino

Ipsum dixit. Il modello black-box come attore epistemico e discorsivo

Il dibattito sull'esistenza e sul ruolo di "black box theories" e più in generale di una "black-box science" all'interno del discourse scientifico ha visto, con forme diverse e diverse accezioni, un'ampia riflessione a partire almeno dagli anni Sessanta (Bunge 1964, Whitley 1972, Latour 1987). L'introduzione, in anni recenti, di modelli di machine learning nel novero degli strumenti utilizzati nella ricerca empirica ha inaugurato un'accezione affatto particolare di questo concetto: restringendo il controllo dell'operatore umano alla sola architettura di alto livello, alla macchina è appaltata la definizione stessa della legge matematica che lega gli input agli output nel dataset di addestramento. Questa operazione fa problema, sotto il profilo epistemico, in particolare quando l'esistenza a priori di una legge siffatta sia ignota o dubbia. Un esempio di particolare rilevanza, non solo per la sua chiarezza ma per le evidenti ricadute etiche e deontologiche, è una riproposizione delle teorie fisiognomiche– e l'apertura di una vera e propria "new era of computational physiognomy" (Stark e Hutson 2022) – in un'area disciplinare a cavallo tra la computer vision e le scienze sociali e psicologiche. È il caso, per esempio, di studi neo-lombrosiani che sostengono di individuare un legame diretto (nei termini appunto di una legge matematica) tra la conformazione facciale e la propensione al crimine. Adottando proprio questo fenomeno come case study, l'intervento si propone di analizzare il ruolo epistemico e retorico che l'AI assume in queste linee di ricerca, indagandone il rilievo nella costruzione e nella legittimazione del sapere (pseudo)scientifico.

*

Ipsum Dixit: The Black-Box Model as an Epistemic and Discursive Actor

The debate on the existence and role of "black-box theories," and more broadly of a "black-box science," within scientific discourse has, in various forms and interpretations, been the subject of extensive reflection at least since the 1960s (Bunge 1964; Whitley 1972; Latour 1987). The recent introduction of machine learning models into the toolkit of empirical research has introduced a new, peculiar meaning to these expressions: by limiting human control to high-level architectural design, the machine itself is entrusted with defining the very mathematical law that links inputs to outputs within the training dataset. This procedure raises epistemic concerns, especially when the a priori existence of such a law is unknown or doubtful.

A particularly salient example, both for its conceptual clarity and its evident ethical implications, is a revival of physiognomic theories – heralding a "new era of computational physiognomy" (Stark and Hutson 2022) – in a disciplinary area straddling computer vision and the social and psychological sciences. A case in point is neo-Lombrosian studies that claim to identify a direct relationship, precisely in the form of a mathematical law, between facial features and the propensity for criminal behavior.

Using this phenomenon as a case study, the present talk aims to analyze the epistemic and rhetorical role that AI assumes in such lines of research, examining its significance in the construction and legitimation of (pseudo)scientific knowledge.

Keywords: black-box theories, computational physiognomy, pseudoscience, machine learning models in reseach, AI discourse.

Luca Pezzini è cultore della materia al Dipartimento di Studi Umanistici dell'Università di Torino. È laureato in Matematica e ha conseguito un dottorato in Sociologia dei processi culturali e comunicativi, con una tesi sull'impiego dello storytelling nella comunicazione scientifica a tema ecologico. I suoi interessi di ricerca, collocati a cavallo tra l'ambito umanistico e quello matematico-scientifico, includono la narratologia computazionale e l'impiego del machine learning nell'analisi dei testi, i waste studies (Monsters of the Waste-land. Materialità e abjection dei rifiuti, E|C 39, 2023; Vittime di scarto. Logiche dell'alterizzazione nelle narrazioni di eco-mafia, in Connessi. Identità culturale e società nella transizione ecosostenibile e nell'interazione uomo-ambiente, a cura di P. Adinolfi e R. Sapino, Bonanno 2024) e la riflessione critica sull'AI (Il sonno dell'AI genera mostri. Filtri convoluzionali e pareidolia algoritmica, in Semiotica dei filtri, a cura di F. Piluso e M. Leone, Aracne 2025; Chaos machine, magick learning: sulle interazioni tra magia del caos e intelligenza artificiale, Sinestesieonline 46, 2025; "Scrivi un poliziesco di montagna in italiano": sui mondi possibili dell'AI generativa, in Fuori dalla giungla d'asfalto. La narrazione di indagine si sposta negli spazi naturali, a cura di L. Pezzini e R. Sapino, Collane Unito, in pubblicazione)

Latha Poonamallee

Parsons School of Design, The New School

Mindful Agents: An Ethical Scaffold for Human Judgment in AI-Mediated Environments

The integration of Artificial Intelligence (AI) into high-stakes decision-making poses profound ethical challenges, notably the erosion of human agency through automation bias, algorithmic authority, and cognitive offloading. While technical AI literacy and top-down ethical guidelines are crucial, they often fail to address the underlying psychological vulnerabilities that lead to uncritical deference, thereby compromising responsible human oversight and accountability. This paper argues that to navigate this dilemma, mindfulness must be reconceptualized from a peripheral wellness practice into a core ethical and cognitive scaffold, essential for preserving human judgment in AI-augmented systems.

We propose a novel conceptual framework positioning mindfulness as a foundational pillar for human-centered AI ethics. It operates through three interdependent pathways that directly counter key ethical risks: (1) Attentional regulation to create the critical cognitive space for ethical deliberation, resisting the automatic, unreflective acceptance of AI outputs; (2) Metacognitive awareness to foster essential epistemic humility, enabling users to recognize the fallibility of both human and machine reasoning and to actively interrogate the values and ethical assumptions embedded in algorithmic recommendations; and (3) Cognitive flexibility to support the adaptive integration of diverse stakeholder perspectives, embodied experience, and contextual wisdom that purely data-driven AI cannot replicate.

The synergy of these pathways cultivates epistemic resilience—the durable capacity to sustain reflective judgment and ethical discernment under technological pressure. This shifts the ethical discourse from a narrow focus on governing AI systems to a more holistic model that equally empowers human agents. We outline practical implications for embedding mindful practices— such as structured reflective pauses

and journaling—into professional training, leadership development, and organizational workflows. This ensures that AI serves as a tool for genuinely augmented intelligence, fostering a culture of critical engagement and shared responsibility. By centering the cultivation of mindful human agency, this framework offers a vital and actionable pathway to realizing a truly ethical and human-centered future for AI.

Keywords: Mindfulness; AI Ethics; Epistemic Resilience; Human-AI Collaboration; Algorithmic Authority

Latha Poonamallee is Professor of Management and Social Innovation at the New School in New York City. She received her PhD in Organizational Behavior from Case Western University, Cleveland, OH.

Fatiha Sabiella
Independent Researcher

Attention Pollution: A Social Pathology in the Age of Generative AI

We cannot do everything, everywhere, all at once. Herbert Simon (1971) warned that "a wealth of information creates a poverty of attention." Half a century later, the rapid rise of generative AI has intensified this condition, transforming the attention economy into an attention crisis. Our information-abundant society has entered a new stage; we are now living amid an 'attention pollution' era. I postulate the term as a cognitive and moral saturation of human perception caused by the flood of AI-generated content. Building on Kevin Eikenberry's earlier notion of attention pollution (2015), this paper redefines it for the generative AI era as a state where our collective attention is contaminated by synthetic, contextless information that exhausts rather than informs. While studies on social media fatigue have examined platform-based overload (Bright et al., 2015) and creator burnout (Kwon et al., 2020), these analyses predate generative AI's widespread diffusion. This paper bridges that gap by addressing what happens when information itself becomes synthetically generated; produced in bulk, detached from human intention, and consumed at inhuman speed. In this environment, we experience social media fatigue because our attentional field has become polluted. Our capacity to discern, rest, and connect is eroded by content that endlessly multiplies without meaningful origin or purpose. Drawing from philosophy of technology and digital epistemology, this paper develops a conceptual analysis of attention as a moral commons, a shared cognitive environment vulnerable to contamination. By linking Simon's economics of attention scarcity (1971) with Byung-Chul Han's "burnout society" (2010), it frames 'attention pollution' as both an ontological transformation and a social pathology of AI-mediated life. Extending the conclusion of Nah et al (2023), the ethics of AI must broaden beyond fairness and regulation toward the stewardship of attention itself, an ecology of focus essential for autonomy, empathy, and democratic digital health.

Keywords: Attention Pollution, Generative AI, Social Media Fatigue, Digital Epistemology, Social Pathology.

Fatiha Sabiella is an independent researcher examining philosophy of technology, digital thanatology, and AI ethics. Her work combines five years of entrepreneurial experience in the tech industry with philosophical inquiry into AI's social and ethical implications, particularly focusing on how AI systems reshape collective knowledge and attention. She currently has two single-authored papers under review for IEEE CAI 2026 and IASEAI 2026.

Luigi Scorzato
Accenture AG

Reliability and Interpretability in Science and Deep Learning: Epistemological Foundations for Responsible AI

This paper examines the reliability of Machine Learning (ML) models, particularly Deep Neural Networks (DNNs), from an epistemological perspective. While DNNs have achieved remarkable success in various domains, their reliability assessment faces fundamental challenges that go beyond standard statistical error analysis.

The research establishes that all models—both traditional scientific (TS) and ML— inevitably depend on assumptions that cannot be fully justified empirically. However, DNNs diPer significantly from TS models in their epistemic complexity: they employ vastly more parameters (that should be seen as cotributing to the complexity of the assumptions) and exhibit greater sensitivity to initialization details, making their assumptions harder to identify and evaluate. The paper introduces the concept of "epistemic complexity" as a language-independent measure of a model's assumptions. This complexity is directly linked to interpretability—defined not as understanding every computational step, but as comprehending the assumptions underlying a model's predictions. Interpretability thus becomes a precondition for reliability assessment rather than merely a desirable feature for explaining AI to non-experts.

The analysis demonstrates that DNNs' high epistemic complexity hinders three crucial aspects of reliability: (1) interpretability of assumptions, (2) definition of state-of-the-art model classes for error estimation, and (3) clear paths for scientific progress. The author suggests potential ways forward, including developing methods to extract measurable features from DNNs, focusing on domains where complete input spaces can be covered, and studying exactly solvable DNN limits. This research contributes significantly to the conference themes by addressing epistemological foundations of AI, providing philosophical frameworks for evaluating AI reliability, and oPering insights for responsible AI development that balances predictive power with interpretability—essential for ethical AI applications in sensitive domains.

Keywords: Reliable AI; Interpretability; complexity; Foundations of statistics

Luigi Scorzato is applied scientist at Accenture. He is a multifaceted professional whose career spans theoretical physics, high-performance computing, philosophy of science, and artificial intelligence. His diverse expertise has allowed him to make significant contributions across multiple disciplines and sectors. Scorzato spent many years as a researcher in theoretical and computational physics, producing over 80 scientific publications. He is well known for the distributed Monte Carlo algorithms that he has proposed to simulate particularly challenging Quantum Field Theories. In recent years, Scorzato has turned his analytical skills toward understanding artificial intelligence. His background in both theoretical physics and philosophy has given him a unique perspective on AI development, capabilities, and implications. He has contributed important insights regarding AI methodologies, interpretability, and the philosophical underpinnings of machine learning systems.

Ludovica Schaerf
Max Planck Society, University of Zurich

Knowledge Representation in Generative Vision Models

In 2023, large image generation models (such as Stable Diffusion and Midjourney) demonstrated to the general public an incredible ability to reproduce visual scenes, deeply embedded into cultural standards and aesthetic motifs. Their exceptionally vast knowledge is stored in the model weights and organized into a series of representational entities internal to the model (latent). This begs the question: what internal representations are created? How can we interpret these representations technically and metaphorically? What do these representations tell us about: the functioning of the models? And its "episteme"?This talk examines the evolving nature of internal representations in generative visual models, focusing on the conceptual and technical shift from GANs and VAEs to diffusion-based architectures. Drawing on Beatrice Fazi's account of synthesis as the amalgamation of distributed representations (Fazi, 2024), we propose a distinction between "synthesis in a strict sense", where a compact latent space wholly determines the generative process, and "synthesis in a broad sense," which characterizes models whose representational labor is distributed across layers. Through close readings of model architectures and a targeted experimental setup that intervenes in layerwise representations, we show how diffusion models fragment the burden of representation and thereby challenge assumptions of unified internal space. By situating these findings within media theoretical frameworks and critically engaging with metaphors such as the latent space and the Platonic Representation Hypothesis, we argue for a reorientation of how generative AI is understood: not as a direct synthesis of content, but as an emergent configuration of specialized processes.

Keywords: image generation models; representation; synthesis; latent space.

Ludovica Schaerf is a PhD student in Digital Visual Studies between the Max Planck Society (MPG) and the University of Zurich (UZH) since March 2023. She holds a Bachelor's in Liberal Arts and Sciences from Amsterdam University College and a Master's of Science in Digital Humanities at the Swiss Federal Institute of Technology of Lausanne (EPFL). Ludovica worked as a Data Scientist for insurance and academic publishing and as an AI developer in the art market. Her interests lie in interdisciplinary research between the Arts, Artificial Intelligence, and Philosophy. Her research focuses on understanding latent spaces of generative vision models from a technical and philosophical perspective.

Cristina Voto
Università degli studi di Torino

From Pixel to Proxies: Toward an Ethics of AI Recognition

This paper develops an ethics of recognition adequate to machine vision by reconstructing how the face becomes a computational proxy within AI infrastructures. Building on a media-archaeological and semiotic method, it articulates a triad of operative praxes—ratio (modeling/syntax), spatio (archiving/semantics), and dispositio (tagging/pragmatics)—through which facial images are reformatted into tokens of operability and made to perform identity. Case studies range from early digitization and XXth mid-century

human–machine experiments in facial measurement to the standardization of test images and contemporary dataset regimes. The analysis tracks how proxies sediment typifications, enact homophilic clustering, and stabilize classificatory decisions that migrate from experimental setups into everyday governance and platformed life. Artistic practices that reverse-engineer recognition serve as a second line of inquiry, exposing the diagrammatic conditions of legibility and opening design space for ethical intervention. The paper advances three normative proposals: infrastructural accountability for archival lineages and dataset semantics; semantic plurality to counter closure through typification; and operable opacity as a right to modulate one's machinic legibility without forfeiting access to services. The result is a framework for human-centred AI that evaluates systems at the level of enunciative operations—how images are modeled, stored, and labeled—rather than at the surface of outputs alone. By treating the facial image as a proxy and a meta-language of computer vision, the paper clarifies the stakes of recognition in AI and offers actionable criteria for assessment across research, policy, and design.

Keywords: facial recognition; proxy; semiotics; ethics of AI; identity

Cristina Voto is a fixed-term researcher (RTD-A) at the University of Turin, where she teaches Digital Design Languages and Semiotics of Cultural Heritage: Perspectives on Intersectionality, Ethics, and AI. She serves on the doctoral boards of Diseño y Creación (Universidad de Caldas, Colombia) and Artes y Tecnoestéticas (Universidad Nacional de Tres de Febrero, Argentina). She is Vice President of FELS — the Latin American Federation of Semiotics. She has collaborated with universities and art institutions including Vrije Universiteit Amsterdam, Universidad Autónoma de Madrid, the University of the West of England, Universidad de Buenos Aires, Universidad Nacional de Colombia, and the Bienal de la Imagen en Movimiento. Her research intersects visual and design semiotics, philosophies of technology, and feminist/queer theory, taking AI art as a field of theoretical experimentation. On these topics, she has published articles, edited volumes, and the monograph Monstruos Audiovisuales. Agentividad, movimiento y morfología (2022).

Xiaotong Li

Department of Philosophy, Sun Yat-sen University

Constructivism and the Binding Problem in AI

This paper explores the binding problem in Artificial Intelligence (AI), with a focus on its significance for achieving more advanced cognitive capabilities. A core prerequisite for the development of AI toward Artificial General Intelligence (AGI) is the possession of human-like cognitive binding capacities. In humans, binding refers to the innate ability to integrate multimodal, and heterogeneous information into structured representations. In AI, the binding problem refers the question of how artificial cognitive systems combine distributed inputs into coherent structures that support high-level cognitive functions such as perception, understanding, and reasoning. Currently, artificial cognitive systems faces a twofold challenge. On the one hand, neural network systems offer powerful distributed computation across massive datasets, but they fall short in representing explicit structures and compositional rules, making it difficult to sustain coherent relational mappings among cognitive constituents. On the other hand, symbolic systems are characterized by well-defined structures and deterministic reasoning mechanisms, which enable abstract cognitive processing. However, their pre-programmed rule sets are too rigid to cope with unexpected or rapidly changing information. On this basis, Emergentism and Structuralist Reductionism have emerged as the two dominant philosophical frameworks in this discourse. According to Emergentism, the properties of a whole are

not reducible to its parts but arise from nonlinear, self-organizing interactions among them. In AI, this perspective suggests that that structures in neural networks are dynamically generated through the coordinated, synchronous activation of neural units, where binding is seen as the emergent result of these patterns. While this model underscores flexibility and generative capacity, it lacks explanation for how such structures remain stable. Structuralist Reductionism argues that the meaning of a whole comes from the arrangement of its parts within a system. It sees binding as connecting cognitive elements through pre-established structural rules, which ensures clarity and interpretability. However, it lacks an account of how these structural rules themselves originate and develop. Thus both dominant philosophical approaches fall short of providing a comprehensive account of the binding problem in artificial intelligence. I argues that Constructivism can serve as a synthetic position. It maintains that cognitive structures are both actively constructed during integrative processes and dynamically self-constraining as they evolve. This claim effectively explains the hierarchical binding mechanism resulting from the integration of neural and symbolic architectures, thereby offering a unified explanatory framework that captures both the generative flexibility and structural clarity of binding.

Keywords: Binding Problems, Artificial Intelligence, Emergentism, Structuralist Reductionism, Constructivism

XIAOTONG LI is a second-year graduate student in the Philosophy Department at Sun Yat-sen University in China. Specializing in the Philosophy of AI and Cognitive Science, their research primarily investigates the "binding problem" in neural networks and the epistemological implications of AI for Science. Recently, She completed a four-month research internship at the Institute of Intelligence Science and Technology, where they focused on Brain-inspired Algorithms and Neural Network Models. Her current work aims to develop epistemological frameworks for understanding AI-led scientific paradigm shifts.

2.
CREATIVITÀ E IMMAGINARIO
CREATIVITY AND IMAGINARY

Contributors:

1. Marinella Belluati
2. Giorgio Busi Rizzi
3. Niccolò Cencetti
4. Claudia Cerulo
5. Matteo Da Pelo
6. Perrine Gaudry
7. Giulia Guarnaccia
8. Silvia Lilli
9. Vittoria Mascellaro
10. Francesca Medaglia
11. Angelo Oddi, Gianmauro Romagna, Riccardo Rasconi, Paola Panarese
12. Daniel Raffini
13. Dimitri Ruggeri
14. Andrea Sartori
15. Oleksandra Vereschak, Lorenzo Porcaro, Emilia Gómez Gutierrez
16. Viviana Vozzo

Marinella Belluati
Università degli studi di Torino

Intelligenza Artificiale e Società: un dialogo interdisciplinare per un immaginario sostenibile

L'Intelligenza Artificiale (IA) costituisce uno dei principali fattori di trasformazione della società contemporanea, incidendo profondamente sulle strutture sociali, culturali e politiche. Sebbene l'attenzione sia rivolta alle innovazioni tecnologiche, risulta fondamentale esaminare le narrazioni pubbliche che accompagnano il suo sviluppo. Tali narrazioni, lungi dall'essere resoconti neutrali, influenzano attivamente la percezione sociale e guidano le scelte politiche, normative e istituzionali. La comunicazione pubblica relativa all'IA assume, pertanto, il ruolo strategico nella definizione dell'immaginario collettivo e nell'orientamento delle politiche in materia di tecnologia. Le narrazioni che emergono nella sfera pubblica celano rilevanti questioni etiche e sociali contribuendo ad influenzare speranze e timori che si concretizzano in azioni pratiche.

In questa prospettiva, il dibattito sull'intelligenza artificiale richiede un approccio interdisciplinare che integri contributi sociologici e comunicativi, considerando la tecnologia non soltanto come elemento tecnico, ma anche come rappresentazione di valori culturali, visioni del mondo e dinamiche di potere. Un esempio fondamentale in questo contesto è l'analisi del quadro normativo europeo dove l'introduzione dell'AI Act evidenzia il delicato equilibrio tra promuovere l'innovazione tecnologica e tutelare i diritti fondamentali. Le regolazioni non sono solo strumenti regolativi, ma dispositivi narrativi che incarnano i valori etici che dovrebbero orientare la progettazione e l'adozione dell'IA e il tema della giustizia algoritmica, in

particolare, solleva interrogativi profondi riguardo alle disuguaglianze sociali amplificate dall'IA. In questo contesto, la proposta di un modello di governance narrativa diventa essenziale. Questo approccio mira a favorire una maggiore partecipazione civica e a promuovere l'inclusività, affinché l'IA possa diventare uno strumento di equità e giustizia sociale. Diventa cruciale sviluppare spazi di dialogo interdisciplinare, dove i cittadini possano partecipare attivamente alla discussione sulla direzione da prendere e la comunicazione pubblica dovrà evolversi per garantire trasparenza e inclusività, promuovendo una cittadinanza digitale consapevole e in grado di contribuire alla costruzione di un futuro tecnologico più equo e sostenibile.

Keywords: Giustizia Algoritmica; Cittadinanza Digitale; Narrative Tecnologiche; Immaginari Sociotecnici; Inclusione digitale

*

Artificial Intelligence and Society: An Interdisciplinary Dialogue for a Sustainable Imaginary

Artificial Intelligence (AI) is one of the main drivers of transformation in contemporary society, profoundly impacting social, cultural, and political structures. While attention is often focused on technological innovations, it is essential to examine the public narratives that accompany their development. These narratives, far from being neutral reports, actively shape social perceptions and influence political, regulatory, and institutional decisions. Public communication regarding AI thus plays a strategic role in defining the collective imagination and guiding technology-related policies.

The narratives that emerge in the public sphere conceal significant ethical and social issues, contributing to the formation of hopes and fears that translate into practical actions. In this perspective, the debate on AI requires an interdisciplinary approach that integrates sociological and communicative contributions, considering technology not only as a technical element but also as a representation of cultural values, worldviews, and power dynamics.

A key example in this context is the analysis of the European regulatory framework, where the introduction of the AI Act highlights the delicate balance between promoting technological innovation and safeguarding fundamental rights. Regulations are not merely regulatory tools but narrative devices that embody the ethical values that should guide AI design and adoption. The issue of algorithmic justice raises profound questions about the social inequalities amplified by AI.

In this context, the proposal for a narrative governance model becomes essential. This approach aims to foster greater civic participation and promote inclusivity, ensuring that AI becomes a tool for equity and social justice. It is crucial to develop spaces for interdisciplinary dialogue, where citizens can actively participate in discussions about the direction to be taken. Public communication must evolve to ensure transparency and inclusivity, promoting a digitally literate citizenship capable of contributing to the creation of a more equitable and sustainable technological future.

Keywords: Algorithmic Justice, Digital Citizenship, Technological Narratives, Sociotechnical Imaginaries, Digital Inclusion

Marinella Belluati is an Associate Professor at the University of Turin, where she teaches Sociology of Communication, Media Analysis, and Communicating Europe at the Department of Culture, Politics, and Society (https://short.do/Jk59Vh). She edited the volume *Femicide: An Analysis Between Reality and Representation* (Rome, Carocci, 2021) and co-authored *Sociology of Communication and Media Environments* (Milan, Pearson, 2023). Her upcoming work includes *Communicating Europe Between Crisis and Transition* (Il Mulino, 2026). She holds the Jean Monnet Chair *Com4T.EU: Communicating Transitions in Europe*

(2023-2027). For several years, she has conducted research on European elections, communication strategies, and the digitalization of the public sphere in Europe. She is currently the President of the *To-Eu European Studies Centre* at the Department of Culture, Politics, and Society and the Deputy Editor of the journal De-Europa. She is a member of the management committee of *CIRSDe* (Interdisciplinary Centre for Research and Studies on Women and Gender). She coordinates *the Regional Antidiscrimination Observatory* (ORA) in Information in Piedmont and has long researched gender representations in media, politics, and journalism. She is currently leading the University of Turin's public project *Public Engagement AI Debating on AI dissemination*, which has produced the *MagIA.news* magazine (https://magia.news/).

Giorgio Busi Rizzi
Universiteit Gent

AI fought the law: images, authorship, copyright, and the (post)digital

This paper examines the implications of the impact of generative AI in image production. The emergence of these hybrid agents, sitting somewhere between tool and co-author, has the potential to subvert the creative economy of visual arts, especially in the more commodified fields (illustration, comics, etc). This landscape has already been profoundly transformed by (post)digital practices, shaped on one side by the strategies and logics of platform capitalism, and on the other by tactics of free sharing, remix, and reuse. In this sense, the potential shift in creative practices introduced by generative AI opens novel opportunities but simultaneously raises pressing concerns about creativity, copyright, and labor. Issues related to authorship, fair compensation and ideological biases embedded in training data intersect with structural economic inequalities in the distribution of these technologies. The recent surge in AI-generated visual works, both in professional and non-professional contexts, has thus sparked widespread controversy. Visual artists have been particularly vocal in criticizing the collection, nature, and accountability of the data used for training.; in response, several legal frameworks are emerging – one example being the European AI Act, which introduces measures to regulate AI-generated art.

This paper will address these issues by discussing the Internet as a massive repository of easily available data and the way stakeholders seek to exploit and monopolize its resources. It will examine how the impact of generative AI entails antithetical effect on artistic labor, encompassing both a democratizing potential and a capacity to disrupt an already fragile economy. It will question the adequacy of copyright as a framework for protecting artists' rights in the context of (post)digital practices and challenges. Ultimately, it will argue for a radical economic and legal paradigm shift in response to a technological paradigm shift that – given the premises of contemporary techno-capitalism – is already occurring.

Keywords: Copyright; (post)digital practices; techno-capitalism; critical theory; artistic labor

Giorgio Busi Rizzi is FWO senior post-doctoral fellow and adjunct professor at Ghent University, teaching the Comics and Graphic Novels course. His current project investigates authorship in post-digital comics; his previous research analyzed nostalgic aesthetics and practices in comics, and experimental digital comics. He holds a PhD in Literary and Cultural Studies with joint supervision by the Universities of Bologna and Leuven. His contributions have appeared in The Journal of Graphic Novels and Comics, Studies in Comics, European Comic Art, Italian Studies, The Cambridge Companion to Comics and The Routledge Handbook of Nostalgia. He is a founding member of the international research group on Italian comics SNIF – Studying 'n' Investigating Fumetti, and member of several international research groups on comics studies.

Niccolò Cencetti
Università di Roma LUMSA

Femminile virtuale: cyber-femminismo e paradigmi di un'estetica artificiale

Nel più ampio contesto della post-modernità, il corpo femminile tenta di ridefinire le zone di confine tra trasformazioni tecnologiche e immaginario gotico post-moderno, così che la proliferazione di soggetti strutturalmente asimmetrici, nonché la decostruzione della funzione di una dialettica binaria si configurano intrinsecamente come spazi di costruzione della soggettività, la cui codifica si paralizza in uno stato di alterazione, in una attitudine teratologica che collabora all'edificazione del cyber-spazio contemporaneo. La sistematizzazione della femminilità nel contesto post-moderno, dunque, si traduce nell'applicazione di nuovi strumenti critici allo studio del canonico e del convenzionale letterario, proprio perché partecipa sia alla disgregazione della soggettività tradizionale, sia alla dissoluzione del sistema-corpo, smantellando la corporeità sociale intesa come reale e dando vita a quel cyber-femminismo che si configura come espressione della trasgressione. Attraverso un'analisi ragionata di elementi postumani e cyber-punk nella narrativa (si pensi, ad esempio, alla produzione letteraria di Nicoletta Vallorani) e nelle realtà virtuali contemporanee, questo studio di prefigge di esaminare le espressioni di un'alterità corporea femminile come luogo di mutazione e ribellione, così da restaurare i paradigmi di un'estetica artificiale, nonché le planimetrie corporee che rivelano le contraddizioni genetico-ontologiche dei binomi natura/tecnologia, organico/meccanico.

Keywords: donne artificiali; cyber-femminismo; fantascienza; alterazioni corporee

*

Virtual Feminine: Cyberfeminism and Paradigms of Artificial Aesthetics

In the broader context of post-modernity, the female body attempts to redefine the zones on the border between technological transformations and post-modern Gothic imagery, so that the proliferation of structurally asymmetric subjects, as well as the deconstruction of the function of a binary dialectics are intrinsically configured as spaces of construction of subjectivity, whose coding becomes paralyzed in an altered state, in a teratological attitude that collaborates to the construction of contemporary cyberspace. Therefore, the systematization of femininity in the post-modern context translates into the application of new critical tools to the study of the canon and of the literary conventional, precisely because it participates both in the disintegration of subjectivity traditional, both to the dissolution of the body-system, dismantling social corporeality understood as real and giving life to cyber-feminism that is configured as an expression of transgression. Through a reasoned analysis of posthuman and cyber-punk elements in fiction (in particular, Nicoletta Vallorani's literary production) and in contemporary virtual realities, this study aims to examine the expressions of the female body as a place of mutation and rebellion, so as to restore the paradigms of an artificial aesthetic, as well as the plans corporeal that reveal the genetic-ontological contradictions of the binomials nature/technology, organic/mechanical.

Keywords: artificial women, cyberfeminism, science fiction, body alterations

Niccolò Cencetti is a PhD student at the University of Rome LUMSA. In his research he has dealt of literary animality and of the hermeneutic lines of Animal Studies applied to literature of the twentieth century, as well as ecocritical and geocritical perspectives on literature. He is the author of an essay on Psyche's female body aesthetics in contemporary literature, published in fourth fascicle of Una / Κοινν. He is co-editor of the volume Officina sui generis (Effigi Edizioni), which hosts his contribution on Leonora Carrington's

bestiaries, as well as the volume Anthropocene 4800: literature, environment, ecocriticism (Milella) and the essay contained therein dedicated to animalities and disaster narratives in Alfred Kubin and Guido Morselli. Attended as secretary and speaker at the International Conference on Anthropocene Studies 4800. Literature, environment: ecocriticism (Florence, May 2023); at the Buzzati Conference fifty years later (Chambéry, September 2023) and at the XXVI AdI Congress (Naples, September 2023). He collaborated with the Italian Encyclopedia for the writing of the entry Luisa Adorno in the Dictionary of Women in Italy 1730-2020 (Works Treccani) and with the Antologia Vieusseux for a review of Literature and Psychoanalysis. It is currently member of the Editorial Committee of the Ellisse scientific series.

Claudia Cerulo
Universitas Mercatorum

«Why should our bodies end at the skin?». Embodiment e AI tra filosofia femminista e letteratura speculativa

Già a partire dagli anni Novanta il pensiero femminista ha indagato i paradigmi epistemologici dell'allora nascente IA (Alison 1998), criticando la progressiva cancellazione del soggetto perpetrata dai sistemi computazionali (Alison 2000). La presunta neutralità dell'IA occulta gerarchie epistemiche che privilegiano prospettive maschili e occidentali: a questa pretesa universalità la riflessione femminista oppone i "saperi situati" (Haraway 1988), secondo cui ogni conoscenza nasce da una prospettiva parziale, localizzata e incarnata. È proprio l'embodiment a costituire il nodo cruciale del dibattito filosofico contemporaneo sull'intelligenza artificiale. Riproducendo un'epistemologia disincarnata, l'IA perpetua quel dualismo cartesiano mente-corpo che il femminismo filosofico ha tentato di decostruire (Grosz 1994; Gatens 1996), spingendo verso la progressiva dissoluzione della dimensione corporea (Hayles 1999; Suchman 2007). Contro questa rimozione operata dai paradigmi tecnoscientifici "petro-sesso-razziali" (Preciado 2019), tuttavia, la speculazione filosofica e la pratica artistica continuano a interrogare criticamente il rapporto tra corpo e tecnologia. In questo contesto la feminist fabulation (Hartman 2008; Braidotti 2013; Barr 1998) emerge come pratica narrativa capace di immaginare un superamento della dicotomia mente-corpo che caratterizza la cultura occidentale e si materializza negli odierni sistemi algoritmici. Attraverso l'analisi di alcuni estratti del romanzo Die Nacht war bleich, die Lichter blinkten (2019) di Emma Braslavsky – ambientato nella Berlino del 2060 e incentrato sulle riflessioni esistenzialiste di Roberta, un'androide del dipartimento investigativo suicidi – il contributo si propone di interrogare le aporie del disembodiment tecnologico alla luce delle riflessioni filosofiche che si interrogano sul rapporto corpo-intelligenza artificiale. Le riflessioni stranianti del "corpo-macchina" della protagonista diventano strumento euristico per riflettere sulle epistemologie situate in rapporto all'IA, mostrando come la narrazione letteraria possa illuminare filosoficamente le contraddizioni del dibattito teorico e restituire dignità speculativa all'immaginario contemporaneo.

Claudia Cerulo è contrattista di ricerca post-doc in Critica Letteraria e Letterature Comparate presso l'Universitas Mercatorum e attualmente lavora a un progetto che indaga i rapporti tra AI e immaginario contemporaneo. Precedentemente è stata assegnista di ricerca presso l'Università degli Studi di Napoli Federico II collaborando al progetto PANIC – Post Apocalyptic Narratives in Italian Culture (2000-2022) durante il quale ha indagato i rapporti tra embodiment e ecocritica in una prospettiva ecofemminista. Ha conseguito un dottorato di ricerca in Letterature Comparate presso l'Università di Bologna (DESE- Doctorat d'études supérieures européennes). I suoi interessi di ricerca comprendono il rapporto tra letteratura e

psicoanalisi, l'ecocritica e la filosofia femminista. Ha pubblicato diversi saggi in riviste nazionali e internazionali. È membro del gruppo di ricerca SnIF (Studying 'n' Investigating Fumetti) e dell'Osservatorio sul Romanzo Contemporaneo. È Professoressa a contratto di Letterature Comparate e Studi Inter Artes presso l'Università degli Studi di Napoli Federico II. La sua monografia Auditory Perception in XXth Century Self-narratives (Bloomsbury 2025) è in corso di stampa. Di prossima pubblicazione è anche la curatela con R. Cinerari del volume RadicAzioni. Corpi, natura e tecnologia (Orthotes 2025).

Matteo Da Pelo
Department of Education, Psychology, Philosophy, Università degli studi di Cagliari

La creatività artificiale oltre l'intelligenza: una prospettiva strumentale

Può la creatività artificiale essere considerata una estensione strumentale della creatività umana? In questo lavoro, dopo aver ricostruito il quadro teorico dei principali modelli di definizione e valutazione della creatività, dalle 4P di Mel Rhodes alle formulazioni di Boden e Sternberg, si analizzeranno le modalità con cui l'IA generativa, nella sua forma più evoluta di large language models (LLMs), produce risultati assimilabili a prodotti creativi, confrontandoli con il meccanismo creativo umano. In questa prospettiva, la creatività artificiale viene definita come meccanismo generativo non intenzionale e non autentico espresso da sistemi non cognitivi. Per cui, sembrerebbe avere la forma di un mezzo di estensione percettiva e cognitiva che, come il telescopio o il microscopio, amplia e trasforma le modalità attraverso cui l'uomo accede al mondo e ne rappresenta la complessità.

Attraverso il confronto tra i possibili meccanismi creativi, il lavoro mostra come le fasi siano le stesse pur in assenza di intenzione e autenticità. La riflessione si estende ai risvolti etici ed epistemologici di tale visione. In opposizione a un approccio competitivo, il contributo propone la prospettiva della Co-Cre-Ai-tion, una collaborazione sinergica tra uomo e macchina capace di coniugare intenzionalità umana e potenza generativa dell'IA. In conclusione, la creatività artificiale, intesa come strumento di mediazione piuttosto che di sostituzione, diventa una nuova forma di "tecnologia percettiva" che ridefinisce i confini dell'immaginazione e apre un fertile campo di ricerca sulla natura plurale della creatività.

*

Can artificial creativity be regarded as an instrumental extension of human creativity? This contribution addresses this question by reconstructing the theoretical framework of the major models for defining and evaluating creativity, from the Four P model to the formulations of Boden and Sternberg. It then analyses how generative artificial intelligence, in its most advanced form represented by Large Language Models (LLMs), produces outputs that can be assimilated to creative products and compares these with the mechanisms underlying human creativity. Within this framework, artificial creativity is defined as a generative process that is neither intentional nor authentic and that is expressed by non-cognitive systems. On this basis, artificial creativity is interpreted as a form of perceptual and cognitive extension that, in a manner comparable to instruments such as the telescope or the microscope, expands and transforms the ways in which humans access reality and represent complexity. While LLMs replicate structural features of creative activity, such as variation, recombination, and transformation, they do so without understanding, intention, or commitment to meaning.

Consequently, creative products generated by LLMs exhibit novelty and usefulness in a functional sense but lack authorship in any substantive cognitive or intentional sense. The analysis further addresses the ethical and epistemological implications of this position. It is argued that treating artificial creativity as au-

tonomous obscures responsibility, authorship, and evaluative authority. In opposition to competitive models that portray artificial creativity as a substitute for human creativity, the paper introduces the perspective of Co-Cre-Ai-tion, understood as a collaborative framework in which human intention and artificial generative capacity are integrated without conceptual conflation. In conclusion, artificial creativity is best understood as a form of mediating technology rather than as a creative agent. As a new type of perceptual and cognitive instrument, it reshapes creative practice by expanding the space of possible ideas while preserving creativity as a fundamentally human capacity.

Keywords: Large Language Models; Intelligenza Artificiale; Creatività; Filosofia dell'Intelligenza Artificiale

Matteo Da Pelo è dottorando di ricerca in Filosofia presso l'Università degli Studi di Cagliari, sotto la supervisione di Pietro Salis e Antonio Lieto. Dopo la laurea triennale in Filosofia e la laurea magistrale in Logica, Filosofia e Storia della Scienza, ha conseguito un master in Filosofia dell'Intelligenza Artificiale e del Digitale, ambito che costituisce ora il suo principale campo di ricerca. I suoi studi si concentrano sui test di valutazione per l'intelligenza artificiale, dal test di Turing ai moderni benchmark, con particolare attenzione al problema dell'overfitting. I suoi altri interessi di ricerca includono inoltre il design cognitivo dei sistemi di IA e l'analisi della creatività artificiale, tema su cui ha proposto una prima definizione teorica in un articolo in corso di pubblicazione su AI & Society.

Perrine Gaudry

Department of French and Italian, Emory University

Toward an Ethics of Ambiguity in AI Image Generation

While most AI-ethics frameworks privilege transparency and accuracy—equating clarity with truth and ambiguity with error—I argue for an ethics of ambiguity for AI. Generative systems have been trained to stabilise uncertainty, yet they operate through probabilistic distributions and could, in principle, sustain a spectrum of possibilities instead of selecting only the most probable one. Ambiguity is one of the forms through which life and representation take shape: the intersex, trans, or disabled bodies that elude medical and social classification; the blurred photograph that hovers between precision and accident. Ambiguity keeps perception open and dialogic: it requires us to interpret, to imagine, and to respond to what we cannot fully master, capacities that ground both creativity and ethical relation.

Through comparative experiments with two generative pipelines—one standard and one ambiguity-positive—I analyse how AI image generation regenerates the ambiguity found in damaged, archival, and medical photographs. The standard pipeline maximises coherence (high CFG scales, convergent denoising), producing neat, legible images that correct irregularities and suppress difference. The ambiguity-positive pipeline, designed with contradictory prompts (e.g., "masculine/feminine portrait," "vulnerability/strength," "precision/accident") and lower CFG values, sustains multiple probabilities at once, generating hybrid anatomies, doubled postures, or faces that blur between ages or sexes.

The analysis identifies three levels of ambiguity—in the image, in the AI system, and in the body—and argues that an ethics of AI ambiguity should treat them not as noise but as creative and critical resources. For instance, when a generated portrait brings together traits coded as different genders—butch, female, male, or even configurations not yet culturally legible—the aim is not to resolve the multiplicity but to perceive it as coexistence: an ethical stance that recognises plurality and a creative practice that expands the field of what can be seen or imagined.

Keywords: Ambiguity; Generative AI; Ethics; Visual culture; Gender representation.

Perrine Gaudry is a researcher in visual culture, photography, and AI ethics, currently Visiting Scholar in the Department of French and Italian at Emory University (USA). Her work explores how contemporary visual systems inherit and reconfigure the classificatory logics of nineteenth-century medical, colonial, and artistic regimes. She develops ambiguity as an ethical and aesthetic principle within AI image-making and visual epistemology. A practicing photographer, her work on queer visibility has been exhibited at the Steffen Thomas Museum of Art, Martha Gault Art Gallery, and Praxis Gallery, and her photographic archives have been donated to the Bishopsgate Institute (London), the Lesbian Herstory Archives (New York), and the Bay Area Lesbian Archives. She holds a PhD from Emory University and has published in Women & Performance, Contemporary French and Francophone Studies, and Rubriques (Aix-Marseille University). Current articles include "Photographic Heterotopias and the Androgynous Imaginary in Pierre Molinier's Work" (revise and resubmit, Third Text) and "Beyond the Frame: Remainders, Poiesis, and the Transformative Power of Photography" (revise and resubmit, Visual Culture Online).

Giulia Guarnaccia

NABA, Università degli studi di Torino

Contro-visioni: pratiche artistiche a confronto nell'era algoritmica

L'arte contemporanea nel contesto globale attuale, assume un ruolo cruciale come dispositivo di resistenza, riappropriazione tecnologica, ma anche di collaborazione e co-esistenza con nuove forme di complessità. La ricerca, prendendo in analisi alcuni casi studi di opere d'arte e progetti collaborativi, traccia tre traiettorie fluide in cui le pratiche artistiche si intersecano con l'IA e con le tecnologie in generale. L'arte attraverso strategie di reverse power, estetiche decoloniali e post-umane e pratiche collaborative socio-educative, decostruisce le epistemologie dominanti, rivela le dimensioni politiche, materiali e cognitive del potere algoritmico e immagina nuove forme di collaborazione. Ben Grosser, Jake Elwes, Trevor Paglen e Kate Crawford sovvertono i dispositivi di sorveglianza, trasformandoli in strumenti di contro-visualità. L'arte di MiMi Ọnụọha mette in discussione la neutralità del dato, mentre le artiste Ameera Kawash e Sofia Crespo propongono prospettive decoloniali e post-umane in grado di immaginare forme di coesistenza tra la vita biologica e quella artificiale. Accanto a queste pratiche, esperienze collaborative come Mapping the Cobalt Supply Chain, centri di ricerca indipendenti come Data & Society e gruppi interdisciplinari come N.i.n.a. (Nè Intelligente nè Artificiale), mostrano come l'attivismo e la pedagogia possano costruire modalità alternative di produzione del sapere, fondate su trasparenza, accessibilità e responsabilità condivisa.

L'arte, anche attraverso l'IA, non si limita a rappresentare le crisi generate dall'automazione, ma agisce come pratica epistemica capace di smascherare le logiche estrattive e di costruire contro-narrazioni etiche e situate. Le contro-visioni qui esplorate non solo perturbano il funzionamento dei sistemi di potere, ma aprono possibilità per un immaginario tecnologico più critico relazionale e collettivo. IA e Creatività: come la produzione e l'analisi dei prodotti artistici (arti visuali, letteratura, cinema, teatro, fumetti e..) si modifica in relazione all'IA.

*

Counter Visions: Artistic Practices in Comparison in the Algorithmic Era

Contemporary art in today's global context plays a crucial role as a device of resistance, technological reappropriation, and also of collaboration and coexistence with new forms of complexity. This research, by examining selected case studies of artworks and collaborative projects, outlines three fluid trajectories in which artistic practices intersect with artificial intelligence and with technologies more broadly. Through strategies of reverse power, decolonial and posthuman aesthetics, and socio-educational collaborative practices, art deconstructs dominant epistemologies, exposes the political, material, and cognitive dimensions of algorithmic power, and imagines new forms of collaboration. Ben Grosser, Jake Elwes, Trevor Paglen, and Kate Crawford subvert surveillance systems by transforming them into instruments of countervisuality. MiMi Ọnụọha's work interrogates the presumed neutrality of data, while artists Ameera Kawash and Sofia Crespo offer decolonial and posthuman perspectives capable of imagining forms of coexistence between biological and artificial life. Alongside these practices, collaborative initiatives such as Mapping the Cobalt Supply Chain, independent research centers such as Data and Society, and interdisciplinary groups such as N.i.n.a. (Neither Intelligent nor Artificial) demonstrate how activism and pedagogy can build alternative modes of knowledge production grounded in transparency, accessibility, and shared responsibility. Art, including through AI, does not merely represent the crises generated by automation but acts as an epistemic practice capable of revealing extractive logics and constructing ethical and situated counternarratives. The counter-visions explored here not only disrupt the operations of power systems but also open possibilities for a more critical, relational, and collective technological imaginary.

Keywords: arte; tecnologie; estetica; conflitto; ia

Giulia Guarnaccia is a visual artist, activist, and independent researcher. She holds a degree in New Art Technologies from the Brera Academy of Fine Arts (2024) and is completing a degree in Visual Arts and Curatorial Studies at NABA. She graduated in the postgraduate programme in Ethics and AI at the University of Turin (2025). Her practice lies at the intersection of art, technologies, and critical theory, with a particular focus on processes of hybridization between the biological and the artificial, post digital aesthetics, and forms of algorithmic resistance. She regularly takes part in national and international exhibitions and artistic research conferences. Among these is the REACT2025 conference, in collaboration with the CultTech Association, where she presented the theoretical contribution titled "Art in Tension: Situated Practices between Algorithmic Resistance and Material Reuse."

Silvia Lilli
Ricercatrice indipendente

Oltre il plagio: pensare i testi generati nella letteratura del futuro

L'arte generativa minaccia particolarmente la tradizionale definizione di letteratura: mai una macchina era stata capace di riprodurre così bene un'attività posta orgogliosamente in analogia con la forma stessa del pensiero. Il problema riguarda il duplice apporto dell'elemento ideativo e di quello esecutivo: dal momento che la realizzazione linguistica viene concepita in rapporto quasi identitario con l'attività creatrice della mente, è difficile privare l'esecutore materiale della catena verbale (la macchina) del diritto di proprietà intellettuale sul testo generato, né è sufficientemente chiaro come l'essere umano che indirizza l'esecuzione possa considerarsi l'autore di un testo se estraneo al processo della sua realizzazione verbale.

Uscire da questa impasse significa ripensare i criteri di definizione dell'opera letteraria, basati fino a oggi – malgrado conclamati decessi autoriali – sulla sostanziale sovrapposizione tra volontà autoriale, messaggio, e realizzazione verbale, almeno nella fase genetica (Smiraglia 2001). Ciò significa ripensare l'opera secondo due possibili direttrici. Da una parte, il testo generato acquisisce lo statuto di testo letterario se si postula la sua indipendenza dall'autore, riconoscendone il senso unitario nell'intrinseca significatività linguistica, e non nell'intenzionalità che vi è dietro (Danesi 2024). Dall'altra, scindendo l'ideazione e l'esecuzione: il testo è salvaguardato in quanto prodotto intenzionale di un autore, designer più che esecutore dell'opera in senso stretto (Floridi 2025).

L'intervento vuole spingere la riflessione ancora oltre: non solo stimolare l'atto di deissi (Bajohr 2024) che equipari le tipologie testuali umana e artificiale in base alla funzionalità per il ricevente (Raffloer e Green 2025), ma trascorrere dal timoroso "cosa sappiamo fare meglio dell'IA?" (destinato a risposte obsolescenti) al "cosa possiamo fare con l'IA che non possiamo fare senza?". Il successo dell'innovazione tecnologica nell'arte non può dipendere, infatti, dall'efficientamento produttivo, ma piuttosto dal valore aggiunto di espressività che motiva e attrae l'essere umano verso lo strumento, per la gamma di possibilità nuove che esso apre (retoriche, narrative, strutturali): valorizzare le specificità delle due scritture, analogica e digitale, impone l'esplorazione coraggiosa, tanto per gli autori che per i critici, di queste inedite possibilità.

Keywords: Authorship; generated literature; intellectual property; reception.

*

Inventing Tomorrow's Literature: Authorship, Collaboration, and AI

While generative art has posed unprecedented challenges to all artistic fields, it is particularly threatening for the traditional definition of literature. For the first time at this level of fluency, a technological device has proven capable of simulating linguistic production, an activity historically theorized as an extension of cognition itself. While computer scientists distinguish between the abstraction ability and symbolic thinking of human being, on one hand, and probabilistic stochastic procedures, on the other, still the quality of the outcomes is enough to undermine the specificity of humans in this specific task.

One possible compensatory strategy could be to enhance the importance of the human prompt in the generative act. However, this is an inadequate theoretical solution to solve the contradiction at the core: since linguistic realization is conceived in an almost identical relationship with the creative activity of the mind, it becomes difficult to deny the material executor of the verbal chain (the machine) a claim to intellectual property, nor is it sufficiently clear how the human being which triggers the execution of the text could be considered its author when extraneous to the process of its verbal realization.

Resolving this impasse requires rethinking the criteria of definition of literary work, traditionally predicated on the coincidence – despite Barthes and Foucault's proclaimed authorial deaths – of authorial intention, semantic content, and linguistic execution (Smiraglia 2001), at least in its genetic phase. This means to rethink the literary work along two possible lines. On one hand, the generated texts would acquire the status of literary text if we postulate its true independence from the author, recognizing its unitary meaning in the text's inner linguistic significance, instead of the intentionality behind its creation (Danesi 2024). On the other hand, by separating the executive and the ideative phase: the text is safeguarded as an intentional product of an author, designer rather than executor of the work in the strict sense (Floridi 2025). This contribution, by exploring the modalities of interaction between authors and generative AI, intends to extend this inquiry. Its aim is not only to stimulate the deictic gesture through which artistic status is conferred (Bajohr 2024) – equating the human and artificial textual typology based on the function for the receiver (Raffloer and Green 2025) – but also to move from the question: "What can we do better than AI?", destined for obsolescence, to the question: "What can we do with AI that we can't do

without?". The success of technological innovation in the art field cannot depend, indeed, on production efficiency, but rather on the added value of expressiveness attracting and motivating human beings towards the instrument, by virtue of the new range of opportunity that it opens (rhetorical, narrative, structural). By foregrounding the cognitive and material specificities of both human and artificial writing, this contribution proposes a reconceptualization of literary authorship as distributed cognitive practice, reframing generative AI as epistemological provocation rather than existential threat to the definition of literature.

Silvia Lilli holds a PhD in Comparative Studies from the University of Rome Tor Vergata, with a dissertation on La lingua sperimentale nelle opere teatrali di Giovanni Testori (2025). In the field of digital humanities, she has explored computational approaches to linguistic creativity ("Creativity, Invention and Linguistic Analysis. A Case Study," Umanistica Digitale, 9, 19, 2025) and LLMs performance ("ChatGPT-4 and Italian Dialects: Assessing Linguistic Competence," Umanistica Digitale, 16, 2023). She presented at several national and international conference (Surprisal as a Lens for Linguistic Creativity, SIG-DLS Workshop, DH2025; Esplorare l'anomalia: uno studio computazionale sullo stile di Horcynus Orca, ADI 2025) and is also a member of the newly established AIUCD Special Interest Group "Osservatorio per l'IA generativa". In the field of contemporary Italian literature, her research focuses on Giovanni Testori and other 20-th century authors (Pasolini, Berto, Flaiano, D'Arrigo), exploring intertextual relations, authorship theory, and intermedial translation.

Vittoria Mascellaro
Accademia di Belli Arti di Napoli

Etica della visione: verità e autorialità nel cinema nell'era dell'intelligenza artificiale

L'introduzione dell'intelligenza artificiale nei processi cinematografici contemporanei impone una riflessione etica sullo statuto dell'immagine, della verità e dell'autorialità. L'IA non produce semplicemente simulacri, ma genera molteplici versioni plausibili del reale, ridefinendo il rapporto tra percezione, fiducia e responsabilità estetica. In questo scenario, la questione non riguarda tanto l'opposizione tra vero e falso, quanto la capacità dei dispositivi visivi di dichiarare la propria costruzione e rendere trasparente il processo di generazione delle immagini. L'intervento analizza come il cinema diventi oggi un terreno privilegiato per esplorare la verità algoritmica: una verità che non deriva dalla somiglianza con il reale, ma dalla coerenza del dispositivo che esplicita criteri, limiti e punti di vista. Attraverso esempi storici di finzione mediale — come i progetti collettivi Luther Blissett e Darko Maver — e pratiche audiovisive contemporanee che integrano immagini sintetiche, archivi e testimonianze, la ricerca mostra come la ridefinizione dell'originalità e della prova visiva interroghi la funzione etica del cinema nell'ecosistema digitale.

In dialogo con le riflessioni di Ruggero Eugeni, Emmanuel Levinas ed Edgar Morin, il contributo propone un'etica dello sguardo algoritmico fondata su trasparenza, tracciabilità e responsabilità condivisa. L'intelligenza artificiale, intesa in prospettiva umano-centrica, non sostituisce la creatività, ma richiede un nuovo patto di fiducia tra autore, spettatore e macchina: un modello in cui la conoscenza visiva diventa un atto relazionale e consapevole, capace di restituire al cinema la sua funzione critica e conoscitiva nella cultura ipermediata contemporanea.

Keywords: Intelligenza artificiale; cinema; verità visiva; autorialità; etica dei media

*

The introduction of artificial intelligence into contemporary filmmaking processes requires an ethical reflection on the status of the image, truth, and authorship. AI does not merely produce simulacra; it generates multiple plausible versions of reality, redefining the relationship between perception, trust, and aesthetic responsibility. In this scenario, the question is not simply the opposition between true and false, but rather the ability of visual devices to disclose their own construction and make the process of image generation transparent.

This contribution examines how cinema has become a privileged ground for exploring algorithmic truth: a form of truth not derived from resemblance to the real but from the internal coherence of a system that makes its criteria, limits, and points of view explicit. Through historical examples of media fiction—such as the collective projects Luther Blissett and Darko Maver—and contemporary audiovisual practices integrating synthetic imagery, archives, and testimonies, the research shows how the redefinition of originality and visual evidence challenges the ethical function of cinema within the digital ecosystem.

In dialogue with the reflections of Ruggero Eugeni, Emmanuel Levinas, and Edgar Morin, this paper proposes an ethics of algorithmic vision grounded in transparency, traceability, and shared responsibility. Artificial intelligence, understood from a human-centric perspective, does not replace creativity but requires a new pact of trust among author, viewer, and machine: a model in which visual knowledge becomes a relational and conscious act, capable of restoring to cinema its critical and epistemic function within today's hypermediated culture.

Keywords: Artificial intelligence, cinema, visual truth, authorship, media ethics

Vittoria Mascellaro (Monza, 1996) is an independent curator and researcher in the fields of visual arts, cinema, and new media. She is pursuing a PhD in Film, Audiovisual Arts, Sound and Media Studies at the Academy of Fine Arts in Naples. Her practice focuses on exploring the intersections between art and artificial intelligence, contributing to the discourse through academic articles, talks, and exhibitions. She is an cultrice della materia in Sociology of Art at the Academy of Fine Arts in Catania and lecturer for the course Introduction to AI and Ethics in Artificial Intelligence at ITSAR Angelo Rizzoli.

Francesca Medaglia
Sapienza Università di Roma

Narrazioni algoritmiche: la serialità complessa tra creatività, AI e etica

L'applicazione dell'Intelligenza Artificiale alla serialità televisiva rappresenta un laboratorio privilegiato per riflettere sulle sfide etiche e culturali legate alla creatività algoritmica, in quanto negli ultimi anni si sono moltiplicati i casi di sceneggiature interamente o parzialmente prodotte da sistemi di IA a partire da esperimenti pionieristici come Sunspring del 2016 (cortometraggio scritto da un algoritmo di deep learning ideato da Ross Goodwin) fino ad arrivare a progetti seriali più recenti sviluppati con modelli generativi come GPT-3 e GPT-4 che hanno portato alla nascita di episodi pilota come Zone Out del 2018, Nothing, Forever del 2022 e di web-series sperimentali diffuse su piattaforme digitali. Parallelamente le major televisive e i servizi di streaming stanno testando l'impiego dell'IA non solo per la scrittura ma anche per l'analisi predittiva delle preferenze degli spettatori con conseguenti ricadute etiche sul rapporto tra creatività, mercato e autonomia artistica. A ciò si aggiungono piattaforme per la creazione di script attraverso prompt poi sviluppati da AI, quali "Showrunner", che ha creato serie animate quali "Exit Valley" e "Kapitol Punishment" attraverso l'uso di Discord. In tal senso l'intervento intende esplorare la dimensione narratologica degli script

e della serialità complessa prodotta dall'AI e al contempo interrogarsi sull'etica delle narrazioni di questo tipo. L'analisi delle prime serie televisive scritte o co-scritte da AI consente di leggere la creatività algoritmica come luogo di tensione fra automatizzazione e responsabilità etica e al tempo stesso diviene occasione per immaginare prospettive innovative nella produzione audiovisiva contemporanea.

Keywords: Serialità; narrazione; AI; etica; narratologia

*

Algorithmic Narratives: Complex Seriality Between Creativity, AI, and Ethics

The application of Artificial Intelligence to television seriality offers a privileged laboratory for reflecting on the ethical and cultural challenges posed by algorithmic creativity. In recent years, fully or partially AI-generated screenplays have multiplied, ranging from pioneering experiments such as Sunspring (2016)—a short film written by a deep learning algorithm created by Ross Goodwin—to more recent serial projects developed with generative models like GPT-3 and GPT-4, which have led to the creation of pilot episodes such as Zone Out (2018), Nothing, Forever (2022), and experimental web series released on digital platforms. At the same time, television studios and streaming services are testing the use of AI not only for screenwriting but also for predictive analysis of viewer preferences, with significant ethical implications for the relationship between creativity, the market, and artistic autonomy. Added to this are platforms designed to generate scripts through prompts developed by AI, such as "Showrunner," which has produced animated series like Exit Valley and Kapitol Punishment through the use of Discord.

This paper aims to explore both the narratological dimension of AI-generated scripts and the concept of complex seriality, and to question the ethics of these emerging forms of storytelling. Analysing the first television series written or co-written by AI allows us to view algorithmic creativity as a site of tension between automation and ethical responsibility, while also offering an opportunity to imagine innovative perspectives for contemporary audiovisual production.

Keywords: Seriality; Storytelling; AI; Ethics; Narratology

Francesca Medaglia is an Assistant Professor in Literary Criticism and Comparative Literature at the Department of Letters and Modern Cultures at Sapienza University of Rome. Her research focuses on authorship, characters, and transmediality. She has published, among other works, five volumes on co-authored and collective writing (La scrittura a quattro mani- 2014; Asimmetrie ibride nella critica di Antonino Contiliano- 2014; Il ritmo dei tempi in Antonino Contiliano- 2014), on the question of authorship (Autore/personaggio: interferenze, complicazioni e scambi di ruolo. Autori e personaggi complessi nella contemporaneità letteraria e transmediale- 2020), and on intermediality and transmediality (Intermedialità diffusa: la narrazione transculturale metamoderna - 2024).

Angelo Oddi
Institute of Cognitive Sciences and Technologies, CNR

Riccardo Rasconi
Institute of Cognitive Sciences and Technologies, CNR

Gianmauro Romagna
Institute of Cognitive Sciences and Technologies, CNR

Paola Panarese
DigiLab Interdepartmental Research Centre, Sapienza Università di Roma

Bias-Aware AI for Cultural Heritage: Detecting Gender and Ethnic Stereotypes through LLMs

This study, conducted as part of the PRIN PNRR project IMAGES (Inclusive Machine Learning Using Art and Culture for Tackling Gender and Ethnicity Stereotypes), analyzes the capabilities of large language models (LLMs) in detecting and understanding gender and ethnic biases in visual and textual cultural heritage metadata. It addresses a double challenge: assessing how LLMs can contribute to identifying bias within digital archives and understanding how they may reproduce forms of representational inequality. The used corpus, drawn from the Italian Ministry of Culture's Central Catalog (MiC) and semantically harmonized through the ArCo ontology, comprises a balanced set of paintings, photographs, and graphic works. Each item is categorized into one of three representational domains, women, gender-diverse groups and ethnic groups, to test model sensitivity to explicit and implicit bias across heterogeneous cultural materials. The research design integrates three complementary tasks: (i) autonomous image description and bias detection, (ii) textual stereotype recognition, and (iii) guided bias identification through structured taxonomies supporting the creation of bias-sensitive metadata. Quantitative performance measures (precision, recall, consistency) are combined with qualitative evaluations by expert annotators, resulting in an interpretive benchmark for assessing reliability, coherence, and fairness.

By framing these experiments within a human-centered ethical perspective, the study aims to connect technical validation with critical reflection on inclusivity, transparency, and accountability in AI-assisted cultural analysis. The results highlight differences in the interpretive depth and contextual awareness of the used LLMs, especially regarding culturally situated or intersectional forms of bias. Beyond technical evaluation, the project contributes to the broader debate on responsible AI by proposing an interdisciplinary methodology that links data auditing, cultural semiotics, and feminist critical theory. Ultimately, the research demonstrates that bias-aware LLMs can serve as both diagnostic tools and ethical mediators, supporting the design of inclusive metadata ecosystems and promoting cultural equity within digital heritage infrastructures.

Keywords: Bias and Fairness, Large Language Models, Cultural Heritage, Inclusive AI

Angelo Oddi is a Senior Researcher at the Institute of Cognitive Sciences and Technologies of the National Research Council of Italy (ISTC-CNR). His work focuses on Artificial Intelligence methods for the automatic and interactive resolution of planning and decision-support problems. He has authored over 90 publications (h-index 28) in journals, books, and international conferences, with applications spanning space exploration, robotics, quantum computing, and cultural heritage.

Gianmauro Romagna is an architect and research fellow at the ISTC-CNR in Rome. He focuses on digital tools to support decision-making processes in the planning, design, and management phases of cultural

and social health interventions, and more generally in complex decision-making contexts in the construction sector. His work falls within the broader context of the digitalization of the construction industry, with a focus on innovation in decision-making and management processes.

Riccardo Rasconi is a senior researcher at ISTC-CNR in Rome. He focuses on Artificial Intelligence (AI) techniques for the automatic solution of scheduling problems and the study of meta-heuristics for solving complex combinatorial problems. He has published more than 50 articles (h-index 19) in journals, books, and international conferences related to AI, with particular reference to the fields of space, robotics, and quantum computing.

Paola Panarese is Full Professor of Sociology of Cultural and Communication Processes at Sapienza University of Rome, where she coordinates the Master's Degree in Gender Studies, Culture and Politics of Media and Communication. She leads the Sapienza research unit of the PRIN PNRR project IMAGES – Inclusive Machine Learning Using Art and Culture for Tackling Gender and Ethnicity Stereotypes and conducts research on media, gender, and algorithmic inequalities.

Daniel Raffini
Sapienza Università di Roma

Conflitti sociali e mediazione tecnologica nell'AI Literature

L'intervento propone un'analisi della pratica dell'AI Literature con l'obiettivo di metterne in luce la dimensione eticamente impegnata e contro-narrativa, intesa come riscrittura critica dell'immaginario dominante. I sistemi di intelligenza artificiale, in particolare quelli data-driven, incorporano e riproducono conflitti sociali strutturali – legati a gerarchie culturali, etniche e di genere – che si inscrivono nel linguaggio e nell'immaginario e possono quindi essere riattivati nei processi di generazione dei contenuti. In questo contesto, scrittrici e scrittori assumono il ruolo di mediatori di tali conflitti sociali e tecnologici, intervenendo sui sistemi di IA attraverso pratiche di interazione attiva, situata e consapevole, che mettono in discussione la presunta neutralità del mezzo. L'uso dell'IA nella produzione letteraria non si configura dunque come semplice support tecnologico, ma come spazio di negoziazione critica e di riscrittura dei bias incorporati nei modelli linguistici, non con il fine di una loro neutralizzazione, ma come esposizione, torsione e problematizzazione attraverso il gesto letterario. L'intervento analizza casi in cui la mediazione umana nel processo creativo diventa strumento di contro-narrazione e riflessione critica sui dispositivi tecnologici, producendo scarti estetici e discorsivi significativi e disallineamenti stilistici. Questo mette l'AI Literature in relazione con categorie centrali degli studi culturali, quali la transculturalità, gli studi decoloniali e i gender e queer studies. Tra i casi presi in esame figurano A Black Story May Contain Sensitive Content di Lillian-Yvonne Bertram, Machine, Unlearning di Li Zilles e Wash Day di Arwa Michelle Mboya.

*

Social Conflicts and Technological Mediation in AI Literature

The paper proposes an analysis of the practice of AI Literature with the aim of foregrounding its ethically engaged and counter-narrative dimension, understood as a critical rewriting of the dominant technological imaginary. Artificial intelligence systems, particularly data-driven ones, incorporate and reproduce structural social conflicts – linked to cultural, ethnical, and gender hierarchies – that are inscribed in language and imaginary and can therefore be reactivated within processes of content generation. In this context,

writers assume the role of mediators of these social and technological conflicts, intervening in AI systems through practices of active, situated, and conscious interaction that challenge the assumed neutrality of the medium. The use of AI in literary production does not function merely as a technological support, but rather as a space of critical negotiation and reworking of the biases embedded in linguistic models, understood not as their neutralization but as their exposure, distortion, and problematization through the literary act. The paper examines cases in which human mediation within the creative process becomes a tool for counter-narration and critical reflection on technological dispositives, producing significant aesthetic and discursive frictions and stylistic misalignments. This situates AI Literature in dialogue with key categories of cultural studies, including transculturality, decolonial studies, and gender and queer studies. The case studies discussed include A Black Story May Contain Sensitive Content by Lillian-Yvonne Bertram, Machine, Unlearning by Li Zilles, and Wash Day by Arwa Michelle Mboya.

Keywords: AI-Literature; counter-narrative; AI creativity

Daniel Raffini is a research fellow at the Department of Computer, Control and Management Engineering at Sapienza University of Rome. He teaches Digital Humanities at the Faculty of Arts and Philosophy. His research focuses on comparative literature, digital humanities, and contemporary Italian literature. He is a member of the executive committee of SIpEIA. He is currently working on the relationship between literature and artificial intelligence, both from a historical-literary perspective and in relation to the theoretical and ethical issues raised by texts generated by artificial intelligence systems.

Dimitri Ruggeri

Le nuove forme della poesia tra etica e intelligenza artificiale

L'intelligenza artificiale sta trasformando radicalmente il panorama della creazione poetica contemporanea, dando vita a nuove forme espressive come la videopoesia generativa e il poetry slam aumentato.

Questo intervento si propone di esplorare le intersezioni tra IA, etica e creatività poetica, analizzando come i linguaggi poetici si stiano ibridando con le tecnologie intelligenti in direzione di un'estetica relazionale, partecipativa e multimediale. Attraverso esempi concreti di sperimentazione, sia individuali che collettive, verrà discusso il ruolo dell'IA come co-autore o ambiente creativo, mettendo in luce le potenzialità ma anche le criticità etiche legate all'integrazione di sistemi generativi nei processi di scrittura, performance e fruizione poetica.

Particolare attenzione sarà data all'ambito della videopoesia, dove l'IA può intervenire nella composizione visiva e sonora, e al poetry slam, dove l'uso dell'intelligenza artificiale apre a riflessioni sulla soggettività, l'autenticità e la connessione tra performer e pubblico. L'intervento adotterà una prospettiva critico-costruttiva, lontana sia da facili entusiasmi sia da determinismi tecnofobici, proponendo una visione dell'IA come strumento di dialogo e ampliamento dei codici espressivi. L'obiettivo è interrogare le condizioni etiche di questa trasformazione, evidenziando come le tecnologie intelligenti possano essere integrate in modo consapevole nella creazione poetica, promuovendo inclusività, pluralità e nuovi immaginari culturali.

Keywords: Intelligenza Artificiale; Creatività Poetica; Etica; Videopoesia; Partecipazione

*

New forms of poetry between ethics and artificial intelligence

Artificial intelligence is radically transforming the landscape of contemporary poetic creation, giving rise to new expressive forms such as generative videopoetry and augmented poetry slam. This presentation aims to explore the intersections between AI, ethics, and poetic creativity, analyzing how poetic languages are hybridizing with intelligent technologies toward a relational, participatory, and multimedia aesthetic.

Through concrete examples of experimentation—both individual and collective—the role of AI as a co-author or creative environment will be examined, highlighting its potential as well as the ethical challenges tied to the integration of generative systems in poetic writing, performance, and reception. Particular attention will be given to videopoetry, where AI can intervene in visual and sound composition, and to poetry slam, where the use of artificial intelligence prompts reflections on subjectivity, authenticity, and the relationship between performer and audience.

The presentation adopts a critical and constructive perspective, steering clear of both facile enthusiasm and technophobic determinism, and proposes a view of AI as a tool for dialogue and the expansion of expressive codes. The goal is to interrogate the ethical conditions of this transformation, emphasizing how intelligent technologies can be consciously integrated into poetic creation to promote inclusivity, plurality, and new cultural imaginaries.

Dimitri Ruggeri is a poet, video poet, slammer, and writer. He is the author of several poetry collections, including Parole di grano (2007), Carnem Levare, il Cammino (2008), Status d'amore (2010), Il Marinaio di Saigon (2013, winner of the Mioesordio Critics' Prize 2014), Soda caustica (2015), Krokodil (2018), and Radon (2019). In 2022, he made his debut as a novelist with Pugni al petto, a work performed while running competitive marathons in Venice, Milan, and Rome. He has created video poems, audiobooks, and theatrical performances in unconventional locations such as abandoned villages, agricultural fields, and disused spaces. One of the pioneers of Poetry Slam in Italy, he conceived Slam[Contem]Poetry, the first Italian portal dedicated to this expressive form. He has taken part in national Poetry Slam tournaments, poetry festivals, and major European video poetry events. His works appear in anthologies and literary websites and have been translated abroad. A guest on RAI (Miss Poesia, 2006), he is the artistic director of the Premio Hombres Itinerante di Videopoesia and develops projects in the field of artificial intelligence.

Andrea Sartori
School of Foreign Studies – Department of Italian, Nankai University Tianjin

Cybernetic Analogy and the Ethics of Governing AI

This paper intervenes in the ethical debate on Artificial Intelligence by revisiting a foundational distinction: analogy versus homology in the relation between brains and machines. Since the late twentieth century, thinkers such as Moravec and Kurzweil have argued that consciousness is structurally equivalent to computation, thereby endorsing a "homology" in which intelligence can be uploaded and preserved in silicon. I contend that such claims are both philosophically untenable and ethically hazardous.

Drawing on the original meaning of cybernetics (Wiener, Ashby), I propose instead that AI should be understood through analogy: living systems and machines share feedback and control dynamics, but they do not share an origin, ontology, or evolutionary history. This distinction has immediate normative consequences: because organisms are embodied, metabolizing, and self-preserving, the "governor" (kybernetes) of AI must be human agents, accountable for directing technology toward the preservation of life.

To illustrate what is at stake, I examine Anil K. Seth's theory of the predictive brain. On the one hand, predictive processing aligns with computational models – Bayesian inference, prediction error minimization – making parts of cognition formally tractable. On the other, Seth's recourse to consciousness as embodied experience and to interoceptive selfhood reveals a domain that exceeds computation. This asymmetry underscores why reducing brain and machine to homology is misguided: biological vulnerability and the struggle for survival remain irreducible to algorithmic processes.

For the ethics of AI, the implication is clear: innovation cannot be left to deterministic narratives of technological merger but requires human-centric governance rooted in responsibility for life. Reclaiming the cybernetic tradition as analogy, not homology, thus offers a conceptual and normative framework for steering AI with both precision and ethical conscience.

Keywords: cybernetics, analogy, homology, consciousness, biology

Bio: Andrea Sartori is Associate Professor of Italian Studies at Nankai University in Tianjin, China. He holds a Ph.D. in Italian Studies from Brown University, along with M.A. degrees from Ca' Foscari University of Venice (Philosophy), the Università Statale di Milano (Digital Humanities), and Florida State University (Italian Studies). His research lies at the intersection of modern Italian literature, philosophy, and the ethics of technology, with a particular focus on the concept of biofiction as a metaphor for exploring life, vulnerability, and the interface between biology and artificial intelligence. Sartori is the author of The Struggle for Life and the Modern Italian Novel, 1859–1925 (2022), which examines the literary reception of Darwinism, and Assaliti 45ocu mille luci del cielo. La cultura della percezione (2023), a critical reflection on perception, culture, and technology. He has also published widely on Pirandello, Italian modernism, and intercultural encounters between Italy and China.

Oleksandra Vereschak
Independent Researcher

Emilia Gómez Gutierrez
Joint Research Center, European Commission

Lorenzo Porcaro
Sapienza Università di Roma

Better Writing about AI: Framework of Cliché Expressions about AI in Written Scientific Communication

Expert communication about AI in the media, entertainment, and research tends to include cliché expressions that may propagate misconceptions about what AI is and its possible societal impact. These, at times, alarmist or technologically deterministic "buzzwords" can set unrealistic expectations about AI for the general and expert public, thus influencing the AI up-take and the agenda for AI governance and research. To tackle the AI "story crisis", previous works mostly focused on supporting experts to avoid AI clichés while designing images. Recognizing the need for more responsible representation of AI, we extend this line of work for another important means of communication- scientific writing. Notably, we investigate the following research questions: What are some common cliché expressions used when writing about AI? What kind of misconceptions about AI may these written cliché expressions propagate? To which extent do the AI experts think these expressions actually propagate misconceptions about AI? To tackle these questions, we first present a framework of written cliché expressions about AI, based on a categorized

distillation of the existing decentralized advice on writing about AI. We then 1/3 run an exploratory evaluation of this framework with 21 multidisciplinary AI experts to determine the degree of their (non-)tolerance to the written cliché expressions about AI and the reasons for the divergence of opinions.

Our findings reveal that cliché expressions attributing human-like intelligence, autonomy or moral reasoning to AI are generally not tolerated for creating misleading myths about AI. However, some cliché terms are more tolerated for their widespread use or lack of clear alternatives. We thus propose 4 ways to further support more responsible written communication about AI for academia, international organizations, and the private sector. By refining AI-related terminology and promoting responsible communication practices, our work contributes to improving public understanding of AI and its societal impact.

Keywords: artificial intelligence; scientific communication; AI narratives; cliché expressions; AI myths

Oleksandra Vereschak is an interdisciplinary researcher whose central interests are narratives about AI and AI regulation, people's perceptions of AI, and AI in education. In her last role as a FIG research associate, under the supervision of Markus Anderljung from GovAI, she has investigated the communication of non-EU AI companies regarding the AI regulation in the EU and their releases of AI models. Prior, she was a research trainee in the HUMAINT (HUman Behavior and MAchine INTelligence in the Digital Transformation) team at the Joint Research Centre (JRC) of the European Commission, working on responsible written communication about AI under supervision of Lorenzo Porcaro and Emilia Gómez. This submission is the result of this project, where Oleksandra categorized different types of misleading sentences in a framework, evaluated it based on the examples found in written scientific communication with a group of experts, and provided a set of recommendations for strengthening responsible scientific communication about AI. Oleksandra obtained her PhD in December 2022 from Sorbonne Université, where she worked under supervision of Gilles Bailly and Baptiste Caramiaux on users' trust in AI in the context of high-risk decision-making.

Lorenzo Porcaro is a research scientist specialized in Recommender Systems, and his main interest is understanding how Artificial Intelligence and Information Technology are affecting people's life. Currently, he is a Marie Skłodowska-Curie Postdoctoral Fellow at Sapienza University of Rome, leading the project "Algorithmic Auditing for Music Discoverability" (AA4MD). More recently, Lorenzo served for almost two years as a Scientific Project Officer at the European Commission's JRC, within the HUMAINT team. There, he conducted research on the trustworthiness and auditing of recommender systems, with a particular focus on the Digital Services Act (DSA) in the context of the European Centre for Algorithmic Transparency (ECAT). From 2018 to 2022, he pursued a PhD at the Music Technology Group within UPF's Department of Information and Communication Technologies, titled "Assessing the Impact of Music Recommendation Diversity on Listeners", under the supervision of Prof. Emilia Gómez and Prof. Carlos Castillo. Lorenzo's doctoral research, awarded with the highest distinction, focused on developing methods to assess the impact of diversity in music recommendations on listeners' behaviors and attitudes, providing empirical evidence of how diversity influences user engagement and perception.

Emilia Gómez Gutierrez is a senior researcher at the JRC of the European Commission where she leads the HUMAINT team, providing technical and scientific support to EU AI policies, notably the AI Act and the Digital Services Act, as part of the European Centre for Algorithmic Transparency. Emiilia's main interest is the impact of AI in human behaviour, specifically AI's impact on jobs, decisions, fundamental rights and children. With the research background in the Music Information Retrieval (MIR) field, Emilia became the first female president of the International Society for Music Information Retrieval, and she is particularly

involved in promoting the role of, and increasing opportunities for, women and in improving diversity of the MIR and AI fields. She received a DEA in Acoustics, Signal Processing and Computer Science applied to Music (ATIAM) at IRCAM, Paris (France) and a PhD in Computer Science and Digital Communication at Universitat Pompeu Fabra in Barcelona, Spain.

Viviana Vozzo
Sapienza Università di Roma

Iris Murdoch and the concept of "Digital Realism"

This proposal aims to present an approach to Iris Murdoch's philosophy and apply it to Digital Ethics and AI, suggesting that building a virtual image, strictly influenced by intersubjectivity and moral vocabulary, is not merely an escape, but could be an attempt to reconnect to the world through imagination, which might be defined as "digital realism".

Indeed, Ethics plays a fundamental role in Iris Murdoch though, placing particular emphasis on the importance of context, in relation to which she focuses on the moral concept of "vision", which, alongside imagination and attention, serves as an exploratory tool for engaging with reality. In this framework language plays a fundamental role in Iris Murdoch's thought: every choice is linguistically oriented, as she stated, «words are where we live as human beings and as moral spiritual agents» (Murdoch, 1972). From a literary perspective, driven by imagination, writers navigate a network of meanings that directly impact reality. Literature's urge towards completeness compensates philosophical language, by developing its moral vocabulary through a process of perfectionism. As "word-users" (Murdoch, 1978) we are constantly immersed in literature that we use to «make interesting forms out of experience» (Ibidem), even before actions are taken.

Today, the pervasiveness of digital technologies transforms the way we express ourselves through ordinary language. According to Murdoch, «we are all story-tellers and in this sense we are all literary artists» (Murdoch, 1978). With our "digital stories" we are shaping our identity and defining our "conceptual life" (Diamond, 2006). However, a virtual-based identity necessarily raises the problem of "reality". The use of digital technology, like all discussions on technique allows us to notice the vulnerability of others in a perspective of care and education "about" and "through" the digital and AI. Any contemporary discussion of moral education as a fundamental aspect of moral realism, from a Murdochian perspective, must necessarily take into account the digital dimension, which provides an opportunity to reflect on the concepts that shape our moral lives.

Key Words: Digital Realism; Perfectionism; Ethics; Moral Education; Ethics ad Literature

Viviana Vozzo is a PhD student in philosophy at Sapienza University of Rome, where she is completing a thesis on moral realism and digital realism in the works of Iris Murdoch. Her research interests include the relationship between ethics and literature as well as the ethics of artificial intelligence.

3.
EDUCAZIONE E SOSTENIBILITÀ
EDUCATION AND SUSTAINABILITY

Contributors:

1. Kelly Arenson
2. Giannangelo Boccuzzi, Alberto Nico, Giancarlo Masi, Flavio Manganello
3. Matteo Bona, Francesco De Pascale, Simone Cuconato, Donato Ferrari
4. Matteo Ciaschi, Daniel Dan
5. Ines Crispini, Aldo Pisano
6. Simone Cuconato, Francesco Scarcello, Roberto Beneduci
7. Lorenzo De Stefano
8. Antonio Luca Donato
9. Marco Emilio, Maria Valentini, Elena Mantoet
10. Heike Felzmann
11. Emanuela Guarcello, Francesca Pileggi
12. Michael Ka-Chi Cheuk
13. Ludovica Marinucci, Marco Annoni, Cinzia Caporale
14. Natalie Nenadic
15. Myrtna Nikolaevna Marangoni Kumov
16. Veronica Punzo
17. Noor Rizvi, Gitanjaly Chhabra
18. Gaia Scarponi, Chiara Mecchia, Simone Teglia, Francesco Pro, Syrine Enneifer, Irene Amerini
19. Alessandro Turano
20. Alessio Vaccari

Kelly Arenson
Duquesne University

Talking Ethics with Machines: Large Language Models as Moral Conversationalists

Many fear that AI will lead to the death of education, especially in the liberal arts, where core educational methods such as intensive writing and close textual analysis are easily outsourceable to machines: advanced large language models (LLMs) can already write papers on any prompt and condense an entire course's readings into a few sentences. However, this paper contends that there is at least one important aspect of liberal arts education that AIs have the potential to enrich rather than eliminate: dialogue. LLMs, such as ChatGPT and Gemini, can prove useful as conversation partners, offering students unlimited opportunities to practice talking about the questions, problems, and topics they are studying. A defining feature of liberal arts education—and one that cannot be made obsolete by AI—is its

emphasis on dialogue: in the classroom, this often takes the form of a "Socratic" question-and-answer exchange, through which students learn to interrogate and develop their knowledge under the guidance of a skilled moderator. This "back and forth" dynamic of live discussion is crucial to fostering critical thinking and engagement, yet it is often difficult to achieve in crowded classrooms, with only one moderator for an

entire group of learners. Here, AI can provide valuable support: LLMs can take on the role of moderators, providing real-time "back and forth," tailored to each student's interests and skill level. The paper explores how this might be especially useful in ethics classes, where, perhaps more than in any other subject area, students come to challenge and refine their viewpoints, often through the use of case studies. I describe how an LLM can serve as an effective discussion partner, helping students assess their personal moral commitments as they explore ethical dilemmas of interest to them.

Keywords: discussion; liberal arts; ethics; education

My background is in ancient Greek philosophy, but my recent work focuses on the philosophy of technology, particularly the ethics of robotics and AI. I have presented several papers on technology ethics, I regularly teach courses in this area (e.g., "Ethics and Technology" and "Are Robots People?"), and I published an online article about the ethics of biocomputing for the American Philosophical Association. I am also a paid AI trainer for a large technology testing platform.

Giannangelo Boccuzzi
Alma Mater Studiorum Università di Bologna
Istituto per le Tecnologie Didattiche, CNR

Giancarlo Masi
Università degli studi di Modena e Reggio Emilia

Alberto Nico
Dipartimento di Giurisprudenza, Università degli studi di Bari "Aldo Moro"

Flavio Manganello
Istituto per le Tecnologie Didattiche, CNR

Epistemic Agency e statuto della matematica nell'IA educativa: uno spunto teorico su giustificazione, affidamento e responsabilità

Il contributo propone una riflessione teorica sull'epistemic agency nell'era dell'intelligenza artificiale in ambito educativo, interpretando la matematica che struttura i modelli come spazio in cui si esercitano capacità epistemiche agentiche di docenti, studenti e istituzioni. A seconda delle cornici assunte e delle teorie in rilievo, la matematica è concepita ora come scoperta di strutture indipendenti, ora come invenzione linguistica e pratica storica che orienta rappresentazioni e obiettivi. Una lettura realista la presenta come tessitura autonoma che i sistemi educativi basati su IA (per esempio nell'assessment automatizzato o nell'AI tutoring) ergo intercetterebbero; una lettura convenzionalista la interpreta invece come costruzione storica e sociale che riflette scelte di valore, finalità didattiche e criteri di successo, propri dei contesti formativi. Si teorizza come tali cornici influenzino la forza giustificativa degli output di sistemi di valutazione, raccomandazione o tutoring, i criteri di affidamento razionale da parte di docenti e studenti e l'attribuzione di agency tra persone, macchine e istituzioni educative. La tesi è duplice: (i) anche quando l'IA individua regolarità stabili nei dati, la condotta del sistema dipende da scelte di rappresentazione e di funzione obiettivo (loss, metriche, soglie) che hanno implicazioni normative per equità, inclusione e riconoscimento dei meriti; (ii) riconoscere tali scelte non conduce al relativismo, ma permette di esercitare capacità epistemiche collettive orientate al confronto pubblico e alla deliberazione responsabile in ambito education. Si propone una distinzione operativa tra vincoli (formali e fisici) e scelte (rappresentazioni e obiettivi), per chiarire quando un output offra buone ragioni d'uso in contesti educativi e quando richieda maggiori oneri di prova e controllo umano. In questa prospettiva, il contributo si offre come uno spunto teorico sul possibile nesso

tra modi di intendere la matematica nei modelli di IA educativa e pratiche di governance, suggerendo che concezioni diverse possano orientare in modo distinto i criteri di giustificazione, affidamento e responsabilità nei sistemi formativi.

Keywords: Epistemic agency; realismo matematico; convenzionalismo; giustificazione; governance dell'IA

*

Epistemic Agency and the Status of Mathematics in AI: Realism, Conventionalism, and Their Implications for Justification, Reliance, and Responsibility

This paper offers a theoretical account of epistemic agency in contemporary artificial intelligence by focusing on the role played by the mathematics and statistics that operationalize learning-based models. In such systems, core design components are articulated in formal terms: representations determine how phenomena are encoded as inputs; objective functions specify what is optimized during training; loss functions shape which errors matter most; evaluation metrics define what counts as success; and decision thresholds (or broader decision rules) translate model outputs into actions. These elements are not merely technical details: they structure the epistemic standing of outputs by fixing, often implicitly, the criteria under which an output may function as a reason for belief or action.

We situate this analysis within a philosophical dispute about the status of mathematics. On a realist reading, mathematics is understood as tracking mind-independent structures, and model performance can be interpreted as evidence of contact with stable regularities. On a conventionalist reading, mathematics is treated as a historically situated language and practice that embeds choices about representation, objectives, and success criteria, thereby shaping what counts as adequate performance and acceptable risk. We argue that adopting one or the other stance can be theorized to affect three interconnected dimensions of AI governance: the justificatory force of outputs (when they provide good reasons for use), the conditions for rational reliance (how and when to trust outputs given their limits), and the attribution of responsibility (how epistemic agency and accountability should be distributed across developers, deployers, and institutional procedures).

To operationalize these claims, we propose an analytic distinction between constraints (formal and physical) and choices (representational and goal-directed). This distinction helps to identify cases in which outputs warrant straightforward reliance and cases that require heightened evidential burdens, independent review, or human oversight. The paper concludes by outlining governance implications: structured documentation of representational and objective choices, transparency about success criteria embedded in losses and metrics, scrutiny of threshold-setting and downstream decision rules, attention to differential impacts, and contestation pathways with effective remedies.

Keywords: Epistemic agency; mathematical realism; conventionalism; objective functions; loss and metrics; AI governance.

Giannangelo Boccuzzi è dottorando (XL ciclo) all'Università di Bologna e ricercatore associato presso il CNR-ITD. Si occupa di etica e diritto dell'IA, epistemologia applicata e governance dei sistemi sociotecnici con focus verticale sul dominio educativo. Ha presentato e pubblicato lavori su explainability, responsabilità e impatti etico-legali dell'assessment automatizzato, con attenzione a disclosure, audit e rimedi. Svolge attività di ricerca e progettazione su pratiche di co-valutazione umano-IA e su criteri operativi per giustificazione, affidamento e responsabilità.

Alberto Nico è dottorando (XXXVII ciclo) presso l'Università di Bari. Avvocato d'impresa (Foro di Bari), DPO certificato ed europrogettista con oltre 10 anni di esperienza in programmi UE (Horizon, Erasmus+, LIFE, CERV). Si occupa di diritto tributario e commerciale, proprietà intellettuale, protezione dei dati e diritto dell'innovazione/IA.

Giancarlo Masi è dottorando (XLI ciclo) all'Università di Modena e Reggio Emilia (UNIMORE) in Learning Science and Digital Technologies. Esperto di progettazione europea, svolge ricerca su AI tutoring e inclusione, con particolare attenzione a modelli di supporto adattivo e accessibilità nei contesti educativi.

Flavio Manganello è primo ricercatore al CNR–Istituto per le Tecnologie Didattiche (Genova) e docente a contratto al DISFOR–Università di Genova. Dottore di ricerca in Scienze dell'Ingegneria (curriculum e-learning), lavora su tecnologie educative, learning analytics e personalizzazione, con pubblicazioni e attività progettuali su data literacy e ambienti di apprendimento digitali.

Matteo Bona
Università degli studi di Torino

Francesco De Pascale
Department of Human and Social Sciences,
eCampus University

Simone Cuconato
Department of Physics,
Università della Calabria

Donato Ferrari
Department of Business and Legal Sciences,
Università della Calabria

Geo-Epistemic Framework in Literary Creativity: a comparative study of human and AI textual constructs

This study examines the epistemological differences between human and AI-generated texts, emphasizing the role of geospatial perception in literary creativity. Human-authored texts are grounded in spatial and experiential markers, whereas AI-generated texts lack such contextual grounding. Epistemologically, we draw on the notion of situation space from situation semantics.

The logical framework, grounded on semantic system (L, ⊩) of situation-based modal logic, supports the idea that human narratives are semantically situated, while AI systems operate through statistical associations devoid of lived spatial experience. To explore these distinctions, the study analyzes sixty-four human-authored and 192 AI-generated texts responding to identical literary prompts were analyzed through NLP techniques — document embedding (BERT), topic modeling, and sentiment analysis — combined with statistical and geospatial methods. Quantitative results show that human texts achieved higher cluster coherence (silhouette = 0.62 vs 0.41), stronger neighborhood preservation (0.85 vs 0.72), and greater intra-cluster semantic similarity (cosine = 0.78 vs 0.61). PCA indicated higher variance explained by the first components in human data (67% vs 48%). Visualization through heatmaps and semantic graphs confirmed richer spatial anchoring and thematic diversity in human narratives, while AI texts displayed tighter but less contextually grounded structures. These findings high-light measurable divergences in cognitive modeling between human and AI writing, contributing to the debate on AI's creative and geospatial limitations.

Keywords: Artificial intelligence; Literary creativity; Geographical epistemology; Humanistic geography; Statistical analysis

Matteo Bona is a scholar of comparative literature and literary geography, with a strong philological background and a pronounced interdisciplinary orientation. His research interests lie at the intersection of literary studies, geocriticism, semiotics, and computational methodologies applied to textual analysis, with particular attention to space, memory, identity, and narrative. His scholarly work focuses on the study of spatial and affective configurations in literary and media texts, as well as on the ways narratives shape the social perception of complex phenomena such as risk, violence, and forms of agency. In recent years, he has developed data-driven approaches to narrative analysis, combining text mining, sentiment analysis, and emotion modelling with humanistic theoretical frameworks. He is the author of a monograph devoted to the concepts of place, borders, and subjectivity, and has co-edited volumes and contributed chapters on humanistic geography, literary creativity, and narrative space, including comparative studies between human-authored and AI-generated texts. His articles have been published in peer-reviewed journals in the fields of semiotics, socio-legal studies, and geosciences, addressing case studies ranging from contemporary literature to media narratives.

Simone Cuconato è ricercatore postdoc in Logica (SSD PHIL-02/A) presso il Dipartimento di Fisica dell'Università della Calabria, dove insegna Matematica e Informatica Applicate alle Scienze della Salute. Inoltre, è visiting lecturer in Logica Matematica presso l'Università di Firenze. Dopo gli studi in Logica presso l'Università Cattolica di Milano, sotto la supervisione di Sergio Galvan, ha conseguito il Dottorato di Ricerca in ICT presso l'Università della Calabria e l'IIT-CNR di Pisa. Nel 2023 ha vinto la Poster Competition durante il primo PhD Day organizzato dall'Università della Calabria e, nel 2024, il Best Paper Prize durante la conferenza internazionale ICANTCI 2024. I suoi interessi Di ricerca includono la teoria della dimostrazione, le logiche modali, l'intelligenza artificiale e la filosofia, la storia e la didattica della matematica.

Matteo Ciaschi
Bologna Territorial Research Area, CNR

Daniel Dan
School of Applied Data Science,
Modul University

AI and the Democratization of Knowledge: Methodological Design for an Inclusive European Educational Project

L'integrazione dell'Intelligenza Artificiale (IA) nei contesti educativi rappresenta una sfida cruciale per promuovere l'inclusione e la partecipazione delle persone con disabilita, configurandosi al contempo come strumento di democratizzazione della conoscenza e di riduzione delle disuguaglianze cognitive, linguistiche e motorie. Il presente contributo intende presentare una ricerca in fase di avvio, finalizzata a esplorare le implicazioni etiche, pedagogiche e sociali dell'utilizzo di sistemi di IA nelle scuole come mezzi di supporto all'apprendimento inclusivo. La ricerca prevede un disegno metodologico comparativo che coinvolgera due istituti scolastici, uno in Italia e uno in Austria, scelti per la loro esperienza pregressa nell'integrazione di tecnologie digitali in ambito educativo. Il protocollo di studio comprende: i) osservazioni etnografiche in classe, per analizzare le pratiche d'uso e le dinamiche relazionali emergenti; ii) interviste semi-strutturate con docenti, studenti e operatori del supporto educativo, finalizzate a comprendere percezioni, competenze e criticita nell'adozione di strumenti di ` IA; iii) analisi qualitativa e quantitativa dei dati raccolti, volta

a individuare pattern di comportamento e livelli di maturita digitale in relazione all'uso dell'IA per l'inclusione. Si sta inoltre lavorando per strutturare il progetto in `una dimensione piu ampia, con l'obiettivo di sviluppare un programma di ricerca-azione europeo. Come partner scientifici sono stati coinvolti la Modul University di Vienna e il CNR – Area Territoriale di Ricerca di Bologna, che contribuiranno allo sviluppo metodologico e all'analisi comparativa dei casi di studio applicativi in Austria e in Italia, nel dettaglio presso la regione Emilia-Romagna. I risultati preliminari, attualmente in fase di raccolta, saranno oggetto di una successiva presentazione e pubblicazione al termine del progetto, con l'obiettivo di proporre linee guida per una IA eticamente sostenibile e pedagogicamente efficace. Tale prospettiva mira a delineare un modello di IA inclusiva e democratica, capace di potenziare la giustizia educativa e di garantire un accesso equo e universale alla conoscenza.

Matteo Ciaschi: Responsabile dell'Area Territoriale della Ricerca del Consiglio Nazionale delle Ricerche (CNR) di Bologna. La sua attivita` e focalizzata sullo sviluppo strategico delle infrastrutture di ricerca, sull'innovazione `digitale e sulla valorizzazione della ricerca pubblica. Ha partecipato a diversi progetti in ambito educativo e tecnologico, con particolare attenzione ai temi dell'inclusione, dell'accessibilita e dell'etica dell'intelligenza artificiale nei contesti formativi. E autore di alcune pubblicazioni scientifiche e ha preso parte, in qualit ` a di relatore, ` a conferenze e seminari nazionali e internazionali. Ha inoltre maturato esperienze di docenza universitaria nei settori della gestione della ricerca e dell'innovazione tecnologica, e attualmente e impegnato nella scrittura e ` nello sviluppo di progetti di ricerca dedicati all'intelligenza artificiale applicata all'educazione.

Daniel Dan: Assistant Professor presso la Modul University Vienna, dove svolge attivita didattica e di ricerca `nell'ambito di Intelligenza Artificiale, Data Science applicata, Education, Marketing e Demografia. Il suo impegno per l'introduzione e lo sviluppo dell'IA nell'educazione si riflette sia attraverso numerose pubblicazioni su metodologie innovative e applicazioni dell'IA nei processi didattici sia nelle sue numerose partecipazioni a conferenze accademiche internazionali dedicate all'intelligenza artificiale e alla formazione. Ha contribuito a progetti di ricerca sulla simulazione della mobilita umana e sull'analisi del sentiment nel settore dell'ospitalit ` a.` E' particolarmente attivo nell'integrare strumenti di IA nei processi educativi. Inoltre, ha assunto ruoli di organizzatore e relatore in eventi di rilievo come la Conference of Computer Science and Intelligent Systems (FedCSIS 2025), l'International Conference on Big Data Analytics (AIED 2025) e numerose Summer School internazionali dedicate all'information retrieval e all'intelligenza artificiale. Il suo profilo multidisciplinare e la partecipazione attiva a workshop e commissioni scientifiche testimoniano un solido impegno a favorire l'innovazione e la qualita` nell'educazione attraverso l'uso dell'intelligenza artificiale, a livello internazionale.

Ines Crispini and Aldo Pisano
Università della Calabria

AI&Ethics Literacy: A Pilot Study

This paper reports the results of a pilot study on AI & Ethics Literacy conducted in Italian secondary schools between 2023 and 2025. The project aimed to assess whether structured education in AI ethics is necessary alongside technical AI instruction and whether such training can be integrated into existing curricula, such as civic education. The study involved 193 students from technical institutes and high schools in southern Italy and consisted of an eight-hour intervention per class based on a dialogical Philosophical

Enquiry (PhiE) methodology. Grounded in the Large Ethics Model (LEM) and informed by the PAIA framework of AI risks (Pervasiveness, Autonomy, Invisibility, Adaptivity), the program combined ethical theory, AI fundamentals, and participatory activities. Data were collected through pre- and post-intervention questionnaires, a content knowledge test, and qualitative responses. Results indicate significant improvements in students' ethical awareness, understanding of AI concepts, and recognition of algorithmic bias, with particularly strong engagement among third-year technical students and first-year high school students. The results highlight the importance of integrating AI ethics into upper secondary school curricula in order to foster the development of a critical and informed citizenship, grounded in a perspective of human–machine conviviality that does not undermine personal autonomy and responsibility.

*

Il contributo presenta i risultati di uno studio pilota sull'AI & Ethics Literacy condotto nelle scuole secondarie italiane tra il 2023 e il 2025. L'obiettivo della ricerca era valutare la necessità di una formazione strutturata sull'etica dell'intelligenza artificiale accanto all'educazione tecnica e verificare la possibilità di integrarla nei curricoli esistenti, in particolare nell'educazione civica. Lo studio ha coinvolto 193 studenti di istituti tecnici e licei del Sud Italia e ha previsto un intervento di otto ore per classe, basato sulla metodologia dialogica della Philosophical Enquiry (PhiE). Il percorso didattico, fondato sul Large Ethics Model (LEM) e sul framework PAIA dei rischi dell'IA (pervasività, autonomia, invisibilità, adattività), integra teoria etica, concetti fondamentali di IA e attività cooperative mediante il metodo dialogico. I dati, raccolti tramite questionari pre e post intervento, un test di conoscenza e risposte qualitative, mostrano un miglioramento significativo nella consapevolezza etica degli studenti, nella comprensione dei concetti di IA e nel riconoscimento dei bias algoritmici. Un'elevata partecipazione è stata riscontrata in particolare tra gli studenti del terzo anno degli istituti tecnici e del primo anno dei licei. I risultati evidenziano l'importanza di integrare l'etica dell'IA nel curriculo delle scuole di secondo grado per formare una cittadinanza critica e consapevole in una prospettiva di convivenza essere-umano macchina che non comprometta l'autonomia della persona.

Keywords: Etica dell'IA; Filosofia morale; PAIA; Large Ethics Model; AI and Ethics literacy.

Aldo Pisano collabora con la cattedra di Bioetica ed etica del digitale e Antropologia Filosofica presso l'UNICAL con la prof.ssa Ines Crispini e attualmente è dottorando in Learning Sciences and Digital Technologies per il settore di ricerca di Filosofia Morale. Si è laureato in Scienze Filosofiche nel 2018 all'UNICAL, svolgendo un lavoro di ricerca tesi in etica e antropologia della tecnica con Carlo Rovelli, all'Università di Marsiglia. È stato visiting research presso la Cattedra Unesco RELIA dell'Università di Nantes e presso la City University of New York. È socio della Società Italiana per l'Etica dell'IA, della Società Filosofica Italiana e della Società Italiana di Filosofia Morale; dal 2021 è membro del Direttivo Nazionale "Inventio" (Filò – Università di Bologna); dal 2024 fa parte del comitato esecutivo SIpEIA. È redattore per MagIA (Università di Torino) e per Ritiri Filosofici. I suoi principali ambiti di ricerca riguardano l'etica dell'Intelligenza Artificiale, l' etica narrativa e la didattica della filosofia.

Simone Cuconato
Dipartimento di Fisica,
Università della Calabria

Roberto Beneduci
Dipartimento di Fisica,
Università della Calabria

Francesco Scarcello
Dipartimento di Ingegneria Informatica,
Modellistica, Elettronica e Sistemistica,
Università della Calabria

La dimostrazione logica nella didattica dell'IA: fondamenti epistemologici e applicazioni pratiche per una visione umano-centrica

La dimostrazione logica è al tempo stesso uno strumento epistemologico e un oggetto didattico complesso. Essa occupa un ruolo centrale nella matematica, nella filosofia e nell'informatica: è la pratica che struttura formalmente il ragionamento, identificando schemi validi di inferenza, e su cui si regge l'intero edificio teorico del sapere scientificamente fondato. La proof theory, in particolare nella declinazione strutturale inaugurata da G. Gentzen, offre un quadro concettuale che permette di chiarire la natura inferenziale delle prove e di proporre percorsi didattici innovativi, assicurando al contempo che l'educazione all'IA rimanga centrata sull'essere umano.

Dal punto di vista teorico, l'uso dei tableaux analitici a blocchi nella notazione a sequenti (TABS), appositamente studiati e applicati alla didattica dell'IA, internalizza gli aspetti semantici tipici dei tableaux in una rigorosa procedura algoritmica di tipo sintattico-combinatorio. La loro natura modulare favorisce una visualizzazione immediata dei passaggi cognitivi rilevanti, unendo solidità formale e semplicità pedagogica. Ciò consente agli studenti di seguire in sequenza il processo inferenziale e di sviluppare consapevolezza metariflessiva sulle strategie adottate. Integrati con strumenti di learning analytics, i TABS permettono anche di monitorare pattern ricorrenti, difficoltà logiche e stili di ragionamento, offrendo un supporto didattico avanzato e human-centered.

Sul piano pratico, questi assunti sono stati applicati nell'indirizzo sperimentale "Cl@ssico Digitale" del Liceo Classico "Gioacchino da Fiore" di Rende (CS), sviluppato in collaborazione con l'Università della Calabria. L'iniziativa introduce la disciplina "Logica e Filosofia dell'IA" nel piano di studi, con l'obiettivo di valorizzare la logica come strumento per sviluppare un pensiero critico-analitico e garantire che l'educazione all'IA rimanga focalizzata sull'essere umano. L'uso dei TABS ha permesso agli studenti di confrontare ragionamento umano e artificiale, favorendo un approccio interdisciplinare tra filosofia, matematica e informatica e sviluppando competenze metacognitive essenziali per un approccio critico all'IA. L'esperienza mostra come la logica possa costituire il nucleo di un nuovo curricolo capace di coniugare formazione classica e scientifico-tecnologica, promuovendo trasparenza, rigore e una visione umano-centrica dell'IA.

Keywords: Logic-based AI; Didattica dell'IA; Teoria della dimostrazione; Human-centered AI

*

Logical Proof in AI Education: Epistemological Foundations and Practical Applications for a Human-Centered Perspective

The logical proof is at once an epistemological tool and a complex didactic object. It plays a central role in mathematics, philosophy, and computer science: it is the practice that formally structures reasoning by identifying valid patterns of inference, and upon which the entire theoretical edifice of scientifically

grounded knowledge is built. Proof theory, particularly in the structural tradition inaugurated by G. Gentzen, provides a conceptual framework that clarifies the inferential nature of proofs and supports innovative educational pathways, while ensuring that AI education remains fundamentally human-centered. From a theoretical standpoint, the use of sequent-like tableaux systems (TABS), specifically designed and applied to AI education, internalizes the semantic features typical of tableaux within a rigorous syntactic-combinatorial algorithmic procedure. Their modular nature enables an immediate visualization of the relevant cognitive steps, combining formal robustness with pedagogical simplicity. This allows students to follow the inferential process step by step and to develop meta-reflective awareness of the strategies they employ. Integrated with learning analytics tools, TABS also make it possible to monitor recurring patterns, logical difficulties, and reasoning styles, providing advanced and genuinely human-centered educational support.

On the practical side, these assumptions have been implemented in the experimental track "Cl@ssico Digitale" at the Liceo Classico "Gioacchino da Fiore" in Rende (CS), developed in collaboration with the University of Calabria. The initiative introduces the subject "Logica e Filosofia dell'IA" into the curriculum, with the aim of valuing logic as a tool for cultivating critical-analytical thinking and ensuring that AI education remains focused on the human dimension. The use of TABS has enabled students to compare human and artificial reasoning, fostering an interdisciplinary dialogue between philosophy, mathematics, and computer science, and developing essential metacognitive skills for a critical approach to AI. The experience shows how logic can serve as the core of a new curriculum capable of uniting classical and scientific-technological education, promoting transparency, rigor, and a human-centered vision of AI.

Keywords: Logic-based AI; AI education; Proof theory; Human-centered AI

Simone Cuconato è ricercatore postdoc in Logica (SSD PHIL-02/A) presso il Dipartimento di Fisica dell'Università della Calabria, dove insegna Matematica e Informatica Applicate alle Scienze della Salute. Inoltre, è visiting lecturer in Logica Matematica presso l'Università di Firenze. Dopo gli studi in Logica presso l'Università Cattolica di Milano, sotto la supervisione di Sergio Galvan, ha conseguito il Dottorato di Ricerca in ICT presso l'Università della Calabria e l'IIT-CNR di Pisa. Nel 2023 ha vinto la Poster Competition durante il primo PhD Day organizzato dall'Università della Calabria e, nel 2024, il Best Paper Prize durante la conferenza internazionale ICANTCI 2024. I suoi interessi Di ricerca includono la teoria della dimostrazione, le logiche modali, l'intelligenza artificiale e la filosofia, la storia e la didattica della matematica.

Francesco Scarcello è Professore Ordinario di Sistemi di Elaborazione delle Informazioni (SSD IINF-05/A) presso l'Università della Calabria. I suoi interessi di ricerca spaziano tra IA, complessità computazionale, teoria dei grafi e degli ipergrafi, soddisfacimento di vincoli, programmazione logica, rappresentazione della conoscenza, ragionamento non-monotono e teoria dei database. Il suo lavoro sulla complessità degli equilibri di Nash ha ricevuto nel 2008 l'IJCAI-JAIR Best Paper Prize, mentre il suo articolo "Hypertree Decompositions and Tractable Queries" ha ricevuto nel 2009 il PODS Alberto O. Mendelzon Test-of-Time Award. È membro dell'editorial board della rivista Artificial Intelligence (Elsevier), per la quale ha anche ricoperto il ruolo di associate editor, fino al 2020, ed è EurAI Fellow e AAIA Fellow.

Roberto Beneduci è Professore Associato di Fisica Matematica (SSD MATH-04/A) presso il Dipartimento di Fisica dell'Università della Calabria e Presidente dell'International Quantum Structures Association (IQSA). È stato anche Visiting Marsico Professor all'Università di Denver (USA) e Research Fellow all'Università di York (UK). Le sue principali attività di ricerca si concentrano: i) sui problemi matematici della meccanica quantistica; e ii) sulla derivazione di un modello quantistico per il trasporto di carica nei semiconduttori.

Lorenzo De Stefano
Università degli studi di Napoli "Federico II"

Preserving Epistemic Agency: A Postphenomenological Approach to AI in Education

This paper addresses the ethical and pedagogical challenges of AI integration in education through a postphenomenological lens, reframing the central question from "Do machines think?" to the more pedagogically salient inquiry: "How do LLMs redistribute cognitive agency in learning environments?". Drawing on Don Ihde's mediation modalities (hermeneutic, alterity, background, quasi-embodiment) and Fasoli's taxonomy of cognitive artifacts (constitutive, complementary, substitutive), we demonstrate that LLMs are not neutral tools but active mediators that co-constitute human-technology-world relations. Our framework reveals how educational technologies operate along a continuum from cognitive enhancement to displacement. Constitutive uses scaffold interpretation without replacing reasoning; complementary uses reduce friction while preserving epistemic agency; substitutive uses perform tasks instead of students, risking technologically induced cognitive diminishment. Building on Vygotsky's theories of social learning and scaffolding, we show how Zone of Proximal Development remains applicable in hybrid educational ecologies where both human and non-human agents mediate cognitive development. We translate this analysis into a pedagogical taxonomy that make AI use tangible and accountable while cultivating epistemic virtues. A case study in philosophy teaching illustrates how assessment must shift from product to process through prompt logs, revision trails, and oral defences. Our central claim is that technological neutrality is untenable: ICT, institutional policies, curricula, and assessment regimes actively configure cognitive ecologies. The paper challenges both prohibitionist and unrestricted adoption stances, discussing instead a structured framework for practices that stabilizes constitutive and complementary uses, preserves non-substitutable spaces for independent reasoning, and embeds transparency across educational processes. This human-centered postphenomenological perspective provides actionable guidance for ethical AI governance, ensuring technological transformation serves cognitive flourishing rather than epistemic erosion.

Keywords: AI; Postphenomenology, Epistemic Agency; Mediation; Educational AI

Lorenzo De Stefano holds a PhD in Philosophical Sciences and collaborates with the Chair of Theoretical Philosophy at the University of Naples "Federico II." He has taught Philosophical Anthropology and Theory and Ethics of Big Data at Federico II, as well as Theoretical Philosophy at the University of Basilicata. In 2025 he obtained the Italian National Scientific Qualification for Associate Professor (ASN) for sector 11/C1 (Theoretical Philosophy) . His publications include the monograph Tra cielo e terra: Eugen Fink e l'interpretazione dei presocratici (FedOA University Press), the edited volume Tecnica e coesistenza. Prospettive antropologiche, fenomenologiche ed etiche (Mimesis, 2024), and numerous articles on Heidegger, Anders, technology, ecology, digital culture and the philosophy of artificial intelligence, published in leading Italian and international journals. Dr De Stefano serves on the editorial board of Mechane: Rivista di filosofia ed antropologia della tecnica and is a member of the Italian Society for Theoretical Philosophy (SIFIT), the European Society for Aesthetics (ESA), the Society for Philosophy and Technology (SPT), the Italian Society of Neural Networks (SIREN), HELMeTO (Higher Education Learning Methodologies and Technologies Online), the Nietzsche Gesellschaft, and the Günther Anders Gesellschaft. He has been a visiting researcher at Technische Universität Dresden, the Husserl Archive of Albert-Ludwigs-Universität Freiburg, the University of Tübingen, and the Eugen Fink Research Center at Johannes Gutenberg-Universität Mainz.

Antonio Luca Donato
Sapienza Università di Roma

Beyond Images: From Deepfakes to Synthetic Violence

Recent discussions on AI-generated harm have used the expression synthetic violence to describe the damage produced by generative technologies, from deepfakes to synthetic media. Yet the term remains largely metaphorical and lacks clear theoretical articulation. This paper proposes to explore its conceptual potential: rather than coining a new notion, it aims to clarify what kind of violence we are facing when harm is no longer represented but generated through artificial synthesis.

Drawing on a Wittgensteinian view of grammar as what expresses the order of meaning within a form of life, and on embodied approaches to sense-making, I suggest that synthetic violence names a transformation in the very grammar of visibility. It refers to the processes through which generative systems detach the expressive surface of the body from lived experience, producing new conditions for recognition – and acknowledgement. Emerging most visibly in gendered forms of online abuse such as non-consensual deepfakes, synthetic violence exposes how generative technologies reproduce and amplify existing asymmetries of power, turning the body into a site of algorithmic control. In this sense, the "synthetic" does not merely describe the technical medium, but the reconfiguration of the conditions under which presence, agency, and vulnerability acquire meaning.

This contribution seeks to clarify and systematize the concept, and opens it to discussion considering possible objections: is synthetic violence truly distinct from symbolic or image-based forms of violence, or does it only intensify them? How can we account for the moral significance of a violence whose conditions are co-constituted by technological systems without attributing moral agency to them? By articulating this concept, the paper aims to contribute to an ethics of artificial intelligence attentive to the embodied, relational, and linguistic dimensions of harm.

Keywords: Ethics; Enactivism; Generative AI; Synthetic Media; Gender-based; violence.

Antonio Luca Donato is a PhD candidate in Philosophy at Sapienza University of Rome, specializing in AI ethics and relational approaches to moral cognition. His research focuses on the intersection of enactive cognitive science, the Wittgensteinian tradition, and technology studies, examining how AI systems reshape conditions of recognition and participatory sense-making.

Marco Emilio
Maria Valentini
Elena Mantoet
IUSVE/Università Pontificia Salesiana

Esercizi di futuro. Innescare cambi di paradigmi epistemici e pedagogici nei contesti scolastici pervasi da IA

Tra i tanti ambiti, la diffusione dell'IA nell'esecuzione dei compiti a casa da parte degli studenti interseca questioni interdisciplinari, ancora parzialmente esplorate empiricamente e teoricamente (Turós et al.,

2025), e rappresenta un contesto privilegiato per comprendere come la pervasività dell'IA richieda cambi di paradigma pedagogici, etici ed epistemici. In esso, infatti, vengono trasformate le motivazioni all'apprendimento degli studenti e modificate le abilità attentive e riflessive (Zhai et al., 2024). Emergono implicazioni sia epistemiche, sia etiche: le nuove capacità culturali (Koslicki & Massin, 2025) concesse dagli artefatti IA scardinano le valenze del sapere scolastico, indebolendo potenzialmente la promozione di competenze riflessive e di virtù epistemiche e sociali (De Caro & Giovanoli, 2025). Tale squilibrio incide sull'agentività docente: se il controllo sulle consegne è impossibile e vengono meno le tradizionali forme di apprendimento significativo, l'insegnante affronta il dilemma pedagogico tra una rinuncia della somministrazione di compiti, in ragione della loro irrilevanza, o il persistere, consapevole dell'impossibilità di verificarne l'utilità. Questa dinamica mina la legittimità epistemica dell'asimmetria educativa di insegnante e scuola, generando potenziali contraddizioni tra accompagnamento e coercizione. Ricercando nuovi metodi di revisione degli apparati concettuali pedagogici, etici ed epistemici dei docenti, la tesi che sosteniamo è che la tradizione foresight dei Futures Studies, congiuntamente con il metodo speculativo di Ross (2016), consenta di affrontare sfide della pervasività dell'IA nei contesti di apprendimento scolastico.

Esploriamo tre caratteristiche dell'approccio: a. la ridefinizione della linea temporale-passato, presente e futuro vengono intrecciati, il presente diventa tempo etico delle scelte, contro il mero presentismo; b. la centralità del co-thinking (Ross, 2016), come combinazione personale e collettiva di costruzione di senso; c. il ruolo trasformativo della progettualità anticipante, che, passando da una visione epistemica a una ontologica di futuro, mira a "lavorare con il futuro" agentivamente (Poli, 2019). Sintetizziamo quindi i risultati significativi per la generazione di modelli e pratiche pedagogiche nei contesti scolastici.

Keywords: Agency, artefatti digitali; funzione docenti; Futures Studies; metodo speculativo.

*

Exercises in the Future. Triggering Epistemic and Pedagogical Paradigm Shifts in School Contexts Pervaded by AI

Among the many domains, the spread of AI in students' homework completion intersects with interdisciplinary issues that are still only partially explored, both empirically and theoretically (Turós et al., 2025), and represents a privileged context for understanding how the pervasiveness of AI demands pedagogical, ethical, and epistemic paradigm shifts.

Within this context, students' motivations for learning are transformed and their attentional and reflective abilities altered (Zhai et al., 2024).

Both epistemic and ethical implications emerge: the new cultural capacities (Koslicki & Massin, 2025) granted by AI artefacts disrupt the value structures of school knowledge, potentially weakening the promotion of reflective competences and of epistemic and social virtues (De Caro & Giovanoli, 2025).

This imbalance affects teachers' agency: if controlling assignments becomes impossible and traditional forms of meaningful learning fade, teachers face the pedagogical dilemma of either relinquishing homework because of its irrelevance or persisting with it, fully aware of the impossibility of verifying its usefulness.

This dynamic undermines the epistemic legitimacy of the educational asymmetry between the teacher and the school, potentially generating contradictions between guidance and coercion.

Seeking new methods for revising teachers' pedagogical, ethical, and epistemic conceptual frameworks, we argue that the foresight tradition of Futures Studies, together with Ross's (2016) speculative method, offers a way to address the challenges posed by the pervasiveness of AI in school learning environments.

We explore three features of this approach:

a. the redefinition of the temporal line—past, present, and future become intertwined, and the present becomes the ethical time of choices, countering mere presentism;

b. the centrality of co-thinking (Ross, 2016) as a personal and collective combination of meaning-making;

c. the transformative role of anticipatory design, which—shifting from an epistemic to an ontological view of the future—aims to "work with the future" agentively (Poli, 2019).

Keywords: Agency; digital artefacts; teachers' role; Futures Studies; speculative method.

Marco Emilio: PhD in Philosophy from UniNE (CH); Assistant Professor in Philosophy at IUSVE/Pontifical Salesian University. His research interests revolve around the notion of collective agency from a social-ontological perspective, applied particularly to digital artefacts in the context of civic technologies and ecosocial risks.

Elena Mantoet: PhD researcher at IUSVE/Pontifical Salesian University (40th PhD cycle in "Learning Sciences and Digital Technologies"). Her current main research areas concern the use of artificial intelligence in instructional design, in which she collaborates on the project "Go – Beyond traditional education" as a member of the research team supporting its investigation and development processes.

Maria Valentini: PhD researcher at the University of Padua (38th PhD cycle in "Learning Sciences and Digital Technologies") and teaching assistant at IUSVE/Pontifical Salesian University. Her current main research areas include: ethical and pedagogical implications and the agency of professionals involved in social robotics interventions in hospital settings; educational and ethical impacts of new technologies in post-media society; narrative theories and practices in educational contexts; epistemology of Futures Studies and futures methods.

Heike Felzmann

School of History and Philosophy & Insight, Centre for Data Analytics, University of Galway

Running to stand still: obsolescence as challenge for meaningful assessment in ethics in the age of GenAI

The advances and easy accessibility of generative AI have brought significant challenges to the field of higher education, especially in the area of the Humanities which has traditionally relied on the combined development of critical thinking and writing skills. The academic essay as meaningful form of assessment in the Humanities appears to have become obsolete, but it remains unclear what to put in its place. Technology enthusiasts have highlighted the potential of AI as versatile tool for facilitating deep learning under carefully curated circumstances where ample space for experimentation and meaningful, shared critical reflection can be provided in the classroom. Academic teachers whose priority is to safeguard the integrity of assessment but who want to integrate AI skills into students' learning experience have harnessed their curiosity to experiment with diverse assessment approaches that combine AI use with elements of skills building and critical reflection while retaining feasibility for the context of mass university. However, for many, this has been a process of rapid cycles of innovation and subsequent obsolescence as GenAI capabilities improved. In this paper, I will outline experiences with these cycles of obsolescence of AI-inclusive

assessment design in the context of teaching ethics to philosophy and computer science students in resource constrained settings. I will link those to the requirements of slowness, friction, discomfort and attentiveness for developing critical thinking and skills building which operate as barriers. Paradoxically, it appears that sustainable meaningful use of AI for learning through the continuing rapid advances of GenAI might require carefully managed 61ndividualized or small group engagement by teachers in students' learning, which is in stark contrast to the frequently promoted promise of easy scalability of AI as tool for meaningful learning.

Keywords: AI; authentic assessment; critical thinking; obsolescence

Heike Felzmann is an associate professor in Ethics at the School of History and Philosophy and the Insight Centre for Data Analytics at the University of Galway, Ireland. She works on ethical issues in information technologies and AI, with experience in various European research projects, and on pedagogical approaches to facilitate information technology professionals' engagement with ethical and societal aspects of their work. She is currently leading a project at the Insight Centre for Data Analytics on the use of communities of inquiry as method for academic learning and public engagement for computer science students and researchers.

Emanuela Guarcello and Francesca Pileggi
Università degli studi di Torino

Formare posture critico-costruttive sull'IA nella scuola primaria: verso un sentimento di responsabilità universale

La proposta intende presentare la ricerca internazionale CAIRE attuata presso il Laboratorio "LIFE" del Dipartimento di Filosofia e Scienze dell'educazione dell'Università di Torino e avviata in *partnership* con la Middlesex University di Londra e la West University di Timișoara. La ricerca riguarda lo studio, la progettazione e la sperimentazione di ambienti virtuali tecno-estetici semi-immersivi per la formazione dei bambini della scuola primaria all'ideazione e all'uso critici e creativi delle tecnologie dell'Intelligenza Artificiale. Uno degli obiettivi principali della ricerca è testare la capacità di questa esperienza semi-immersiva di incrementare la competenza riflessiva dei bambini sui limiti e potenzialità dell'IA e la competenza deliberativa di "principi di azione" per una loro "invenzione" e gestione in grado di ampliare la qualità della vita delle creature umane, non-umane e della Terra. Queste competenze (afferenti all'ambito della responsabilità universale) sono state testate utilizzando un approccio di ricerca-azione partecipativa e strumenti di ricerca qualitativa. In particolare, i contenuti emersi sul tema dell'IA (tramite la somministrazione di questionari pre/post) sono stati analizzati attraverso il metodo della *thematic analysis* articolato in sei fasi: familiarizzazione con i dati; generazione delle etichette di testo; articolazione dei possibili temi; revisione e generazione delle mappe tematiche; definizione e denominazione dei temi e, infine, scrittura del report. Dall'analisi è emerso un miglioramento su tre livelli: le conoscenze in merito alla logica di funzionamento dell'IA integrata negli oggetti tecnologici; la comprensione degli errori concettuali e di alcune delle principali misconcezioni antropomorfiche sull'IA e la competenza di problematizzazione. Quest'ultima è stata esercitata nella riflessione sui danni per gli esseri umani, non-umani e per la Terra che scaturiscono da un uso irresponsabile dell'IA e sui benefici di un'ideazione e gestione critica e responsabile dell'IA. Tale lavoro ha condotto all'elaborazione di principi etici orientati alla cura della qualità dell'esistenza delle generazioni future e dell'intera Terra.

Keywords: Esperienza tecno-estetica; Ambienti semi-immersivi; Scuola primaria; Competenza riflessivo-deliberativa; Thematic analysis

*

Developing Critical-Constructive Postures on AI in Primary School: Towards a Feeling of Universal Responsibility

This proposal aims to present the international CAIRE research project, carried out at the "LIFE" Laboratory of the Department of Philosophy and Education Sciences at the University of Turin, launched in partnership with Middlesex University London and West University of Timișoara. The research concerns the study, design, and testing of semi-immersive techno-aesthetic virtual environments for training primary school children to develop critical and creative approaches to the design and use of Artificial Intelligence technologies.

One of the main objectives of the project is to test the ability of this semi-immersive experience to enhance children's reflective competence regarding the limits and potential of AI, as well as their deliberative competence in formulating "principles of action" for its "invention" and management—principles capable of improving the quality of life of human and non-human creatures and of the Earth. These competences (related to the domain of universal responsibility) were tested using a participatory action-research approach and qualitative research tools.

Specifically, the content that emerged on the topic of AI (through pre/post questionnaires) was analyzed using the six-phase method of thematic analysis: familiarization with the data; generation of text labels; development of potential themes; revision and creation of thematic maps; definition and naming of themes; and finally, report writing. The analysis revealed improvements on three levels: knowledge regarding the functioning logic of AI integrated into technological devices; understanding of conceptual errors and major anthropomorphic misconceptions about AI; and competence in problematization. The latter was exercised through reflection on the harms resulting from irresponsible uses of AI for human and non-human creatures and for the Earth, as well as on the benefits that a critical and responsible design and management of AI may bring to social life. This work led to the formulation of ethical principles oriented toward caring for the quality of life of future generations and of the entire Earth.

Keywords: Techno-aesthetic experience; Semi-immersive environments; Primary school; Reflective-deliberative competence; Thematic analysis

Emanuela Guarcello is a faculty member in the field of General Pedagogy at the Department of Philosophy and Education Sciences of the University of Turin and serves as Director of the "LIFE" Lab (Laboratory of Innovation in Philosophy and Education Sciences), which is based in the same Department. Her main research interests include the relationship between technological innovation, Artificial Intelligence, and human development, as well as the relationship between judgment, character skills, and educational processes in both school and out-of-school contexts.

Michael Ka-chi Cheuk
Hong Kong Metropolitan University

Hybrid Intelligence in Higher Education: *Teach* as Case Study of Ethical and Human-Centric Use of AI

The rapid integration of generative AI in higher education raises urgent ethical challenges, including the undermining of academic integrity and the heightened evaporation of teaching-related jobs. While AI tools increasingly automate content creation, the fundamental issues of AI hallucination and bias remain unresolved. Moreover, PhD graduates and experienced scholars face further underutilization in an already cut-throat academic job market.

This paper discusses *Teach*, an EdTech startup project that exemplifies a hybrid-intelligence model combining the efficiency of generative AI with the judgment of PhD-trained editors to produce reliable and readable teaching primers for niche university courses.

Shortlisted for St Francis University's Techno-Humanities Entrepreneurship Competition 2025, *Teach* addresses key ethical concerns such as hallucination and bias through structured human expert oversight. Beyond improving the reliability of AI-generated content, *Teach* introduces a sustainable model of academic labor by creating new roles for underemployed scholars, while saving full-time instructors substantial course-preparation time.

Drawing on findings from the platform's prototype, the paper explores how *Teach* ethically balances technological innovation with human judgment, fostering both AI literacy and institutional trust.

Keywords: AI ethics; hybrid intelligence; higher education; human-centric innovation; academic labor

Michael Ka-chi Cheuk is Senior Lecturer at Hong Kong Metropolitan University and the Founder/CEO of *Teach*. His research and entrepreneurial work focus on ethical AI applications in education, hybrid intelligence, and the future of academic labor. He is also the host of the Spotify podcast *The Cultural Life of the Nobel Prize in Literature Podcast*.

Ludovica Marinucci, Marco Annoni and Cinzia Caporale
CID Ethics-CNR

Research Ethics by Design nello sviluppo di sistemi di Intelligenza Artificiale per l'educazione: il progetto EYE-TEACH

Il progetto europeo EYE-TEACH mira a sviluppare un sistema di Intelligenza Artificiale (IA) basato su dati di tracciamento oculare (Eye tracking) che supporti i docenti nel migliorare le capacità di comprensione della lettura di studenti delle scuole primarie e secondarie. In tale contesto, il Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID Ethics) del Consiglio Nazionale delle Ricerche (CNR) contribuisce all'integrazione di un approccio di Research Ethics by Design, volto a garantire che le dimensioni etiche e sociali siano considerate sin dalle fasi iniziali delle attività di ricerca e per l'intero processo di sviluppo del sistema. L'attività del CNR si concentra sull'analisi dei profili etici e sulla valutazione dell'impatto individuale e sociale di tale tecnologia progettata per l'ambito educativo, con particolare attenzione alla tutela degli interessi, dei diritti e della libertà fondamentali dei partecipanti alle attività di ricerca di carattere empirico

e sperimentale condotte nell'ambito del progetto. Particolare rilievo assume la protezione degli studenti minorenni, inclusi nella fase di valutazione dell'accuratezza del sistema, in conformità con i principi etici – tra cui il rispetto dell'autonomia, la responsabilità, la trasparenza, l'equità, ecc. – descritti nelle raccomandazioni e documenti internazionali per lo sviluppo di un'IA affidabile, nonché nelle linee guida in materia di etica e integrità nella ricerca elaborate dalla Commissione per l'Etica e l'Integrità nella Ricerca del CNR (Commissione CNR) . Questo approccio proattivo non si limita a garantire una condotta eticamente sostenibile del progetto in corso, ma mira anche a produrre raccomandazioni e linee guida utili ai ricercatori per lo sviluppo responsabile di sistemi di IA in ambito educativo, contribuendo a delineare un quadro di riferimento per l'innovazione tecnologica orientata ai valori umani, alla tutela dei partecipanti alle ricerche, in particolare minori, e al rafforzamento della fiducia nell'uso dell'IA per scopi educativi.

Keywords: Intelligenza Artificiale; Etica della ricerca; Ethics by Design; Eye tracking; Tecnologie educative

Ludovica Marinucci is Head of the Ethics and Artificial Intelligence Research Unit at the Interdepartmental Centre for Ethics and Integrity in Research (CID Ethics) of the Italian National Research Council (CNR). Her research focuses on the analysis of the ethical issues of emerging technologies and on the assessment of their societal impact, with the aim of developing recommendations, codes, guidelines, and new ethical paradigms for the design of ethically sustainable AI and robotic systems, particularly within the European projects EYE-TEACH and PILLAR-Robots. She is responsible for the Scientific Secretariat of the Working Group Ethics of Research in Robotics of the CNR Research Ethics and Integrity Committee. She is also an adjunct professor at the University of Rome Tor Vergata and at the International University UNINETTUNO, where she teaches courses on the history and philosophy of technology, philosophy of computer science, and AI ethics.

Marco Annoni: Responsabile dell'Unità di Ricerca "Bioetica" del CID Ethics-CNR e membro della Segreteria scientifica della Commissione CNR.

Cinzia Caporale: Coordinatrice del CID Ethics-CNR e della Commissione CNR.

Natalie Nenadic
University of Kentucky

AI Deepfake Pornography, Sex-Specific Harms, and Law

We are experiencing an explosion of AI deepfake pornography, 99% of whose victims are female, a crisis that raises pressing ethical questions. They center on the human rights of women, especially psychological and other harms, and on implications for human relationships. Attaining a deeper understanding of these harms is imperative to guiding effective regulation where currently there is none and in countering social polarization. I contribute to this understanding by first introducing what AI deepfake pornography is. Here, the perpetrator takes any image of his target and inserts it into an AI app, which integrates that image into mainstream pornography. Her likeness then becomes someone to whom every imaginable sexual abuse is done. The principal victims are teenaged girls targeted by male classmates and women in positions of authority such as politicians and celebrities, the aim being to dehumanize, humiliate, and silence them. Survivors describe the experience as life-shattering. A result is trauma and terror, often resulting in withdrawal from social and public life, including through suicidal ideation and attempts, which effectuates a

kind of second-class citizenship. I then clarify the overwhelming content and effects of mainstream pornography in its earlier technological forms, into which AI integrates these images, and I show how this new technological combination expands a dragnet of sex-specific harms. This new experience-driven (hermeneutical-phenomenological) understanding of earlier technological pornography is the result of decades of feminist research, which formed a foundation that then guided a novel civil rights legal framework for addressing harms. I consider how creative interpretation of this feminist understanding in terms of AI deepfake pornography may guide regulation, including assessment of a patchwork of regulatory proposals. All of this will help mitigate the increasingly aggressive and normalized social polarization across gender/sex to which this crisis is a significant contributor.

Keywords: deepfake pornography; harms; feminism; law

Dr. Natalie Nenadic (Ph.D. Yale) is an associate professor of philosophy at the University of Kentucky, USA. She has published across topics in the philosophy of technology, including on Hannah Arendt, Martin Heidegger, and sexual abuse through Internet-age pornography. Her current research extends this work to feminist analysis and regulation of AI deepfake pornography. She has been awarded funding to conduct a workshop on this topic in Spring 2026. This event follows her Spring 2025 funded workshop "Towards a Philosophy of AI: Life and the Philosophical Canon."

Myrtna Nikolaevna Marangoni Kumov
Universidad Complutense de Madrid

Chromosome AI: Deepfakes and Symbolic Violence Against People with Down Syndrome

This paper critically analyses the use of generative artificial intelligence platforms for creating fake images (deepfakes) depicting people with Down syndrome, particularly in eroticized contexts and with widespread dissemination on social media. Through a mixed-methods approach, including experimentation with AI models, documentary analysis of the European legal framework, surveys with families, and interviews with associations; the study identifies regulatory gaps, ethical challenges, and the lack of participation of affected communities in public debates. The study highlights the technical and economic ease of generating and spreading such content, the limited protection available on social media, and the need to strengthen public policies. It proposes a multidisciplinary approach that combines regulation, digital literacy, and social communication as pillars to prevent abuse and promote an ethical, inclusive, and human rights–oriented use of artificial intelligence.

Keywords: artificial intelligence; deepfakes; Down syndrome; digital ethics; symbolic violence

Myrtna Marangoni is a marketing and communication professional with over 15 years of experience in the technology sector and a Master's degree in Artificial Intelligence for Communication and Media from Universidad Complutense de Madrid. Her academic work explores the ethical implications of AI in representation and digital identity, focusing on the social impact of emerging technologies. With a background from MIT Professional Education and Harvard Extension School, her research bridges innovation, inclusion, and intercultural communication. She is also dedicated to mentoring and developing diversity-focused initiatives that promote equal opportunities and representation in the digital economy.

Veronica Punzo
Scuola Superiore Sant'Anna

AI Literacy within the school context: Legal, Ethical, and Pedagogical Pathways toward Inclusive Education

The use of artificial intelligence (AI) in education (AIED) has revealed significant potential, particularly for innovating teaching and learning practices in the context of achieving SDG 4, which aims to ensure inclusive and equitable quality education and promote opportunities for lifelong learning for all. Rapid tech advances create risks, including insufficient training and lagging adaptation in education and society. Integrating AI in schools also raises ethical, social justice, and child rights concerns. Automated decision-making in teaching and assessment can introduce biases, profiling, and discrimination. Minors remain legally vulnerable, even before they become digital users. Through GDPR, the AI Act, and the Digital Services Act, Europe provides strong protections for minors' data and digital rights. These regulations set age limits for consent, require transparency, prohibit exploitative AI, and classify educational AI use as high risk, mandating strict oversight. Platforms must assess and mitigate risks for young users. Legal safeguards alone are not enough: ethical design, pedagogical responsibility, and digital literacy for students, teachers, and families are crucial. Introducing AI in schools requires a holistic approach that goes beyond simple legal protection, integrating ethical design principles and strong pedagogical responsibility. he AI Act supports educational AI but stresses that schools must guide its critical and responsible use for students' benefit. In an attempt to develop a develop a framework that aims to create synergies among stakeholders in the educational ecosystem (school leaders, teachers, families, and students) that leads to the construction of AI literacy in the educational context, we refer to the Index for Inclusion. This framework encourages the participatory processes [1] with the aim of promoting practices that ensure legal security, foster pedagogical innovation, and maintain technological sustainability, as well as promoting digital well-being in line with ongoing cultural and technological transformations.

Key words: AI Literacy, Legal and Ethical Responsibility, Inclusive Education

Veronica Punzo is a PhD candidate in the National PhD in Artificial Intelligence for Society at the University of Pisa and a member of the LIDER-Lab at the Scuola Superiore Sant'Anna. With an interdisciplinary background in law and pedagogy, her work focuses on the legal, ethical, and social implications of equity, inclusion, and diversity in education, with particular attention to the impact of emerging digital and AI-based technologies. She is an attorney at the Macerata Bar and a secondary upper school teacher of economic and legal sciences with a specialisation in special educational needs support. Her expertise lies at the intersection of accessibility, fundamental rights, and educational governance. Her research focusses on the regulatory aspects of AI-based technologies, with a particular emphasis on the protection of minors and vulnerable students. Her goal is to develop regulatory and operational frameworks based on fundamental rights protection and inclusion to support the responsible adoption of AI in education. She is actively involved in research initiatives at the European and international levels and has received academic mobility grants, including a DAAD scholarship. As part of these projects, she works on public policy analysis, research protocol development, and institutional feedback on European Union regulatory instruments.

Noor Rizvi
Kansas State University

Gitanjaly Chhabra
University Canada West

Digital Inclusion as Ethical Imperative: Reimagining Access and Participation in AI-Mediated Education

The swift uptake of artificial intelligence (AI) by the education sector has transformed knowledge transmission, access, and assessment. Although AI provides personalized learning and increased access, it typically depends on an assumption of worldwide digital capability and connectivity conditions that might not hold for all. This paper discusses that digital inclusion should be regarded as an ethical imperative, not a second-tier concern, within the development and implementation of AI-mediated education. Without deliberate equity consideration, AI technologies can perpetuate digital divides as well as replicate structure-based inequities with respect to culture, language, and economic status.

Drawing on Sen (1974) and Walsh (2003)'s capability approach, Ficker's (2007) hermeneutical injustice, arguments on AI for education (Sinha, 2025; Selwayn, 2024), and digital ethics paradigms, this paper addresses the ethical and psychological dimensions of inclusion in AI-enhanced learning environments. It addresses two research questions guiding the study: (1) How do linguistic, cultural, and socioeconomic contexts frame the human-AI learning relationship? and (2) What are the ethical theories that can be used to guide inclusive and equitable AI design in education?

Through critical integration of new research and policy reflection, such as UNESCO's Recommendation on the Ethics of AI and EU recommendations on responsible AI, the paper sketches opportunities and issues for inclusive AI design. It proposes design principles based on ethical necessities such as transparency, co-creation with users from various backgrounds, and flexibility to specific contexts to make AI tools enable equitable participation and not hinder it. Overall, this paper reframes digital inclusion as an existential dimension of pedagogic ethics in the age of AI. By merging ethics, psychology, and policy, it urges us to rethink access and participation such that technology is a bridge, and not a boulder, to educational justice.

Keywords: AI-mediated education; digital inclusion; human-AI learning relationship; inclusive AI

Dr. Noor Rizvi holds a Ph.D. in English Language Teaching and is a Fulbright Scholar (The University of Kansas, USA, 2018–2019), where she also served as a Cultural Ambassador of India. She has published around ten research papers, including book chapters, and has presented them at major international conferences, including Seattle University and Yale University. Her research interests include Applied Linguistics, AI and Education, Moral Imagination in Healthcare, Indigenous Studies, Digital Media in English Teaching, and English as a Second Language (ESL). Dr. Rizvi currently teaches Writing Skills in the Department of English at Kansas State University and teaches as part-time faculty at Manhattan Technical College, USA. She is additionally trained in mental health therapy and certified in CPR.

Gitanjaly Chhabra is an Assistant Professor at University Canada West (UCW), Vancouver, Canada. She is an academic, philosopher, and writer. She has published articles in prestigious journals and presented her work at prominent conferences. Her current research work focuses on philosophy of artificial intelligence (AI), posthuman philosophy (consciousness and applied linguistics), technology mediation, decolonization and green education. She believes in the 'phenomenological immersion in infinite unity' - exploring the boundaries of self and beyond. She loves coffee and travelling.

Gaia Scarponi, Chiara Mecchia, Simone Teglia, Francesco Pro, Syrine Enneifer and Irene Amerini
Department of Computer, Control and Management Engineering "A. Ruberti",
Sapienza Università di Roma

Teaching Machines Fairness: Data Poisoning for Bias Reduction in AI Models

Web-sourced datasets inherently reflect societal inequalities and stereotypes related to gender, ethnicity, and other social categories. Consequently, Artificial Intelligence (AI) models trained on such large-scale, web-scraped data inevitably encode and reproduce these biases. This can manifest in their internal representations and outputs and may exhibit discriminatory behaviours towards specic groups based on sensitive attributes such as gender and race. Specifically, we quantify these biases by assessing the tendency of Vision–Language Models (VLMs), particularly CLIP, to associate certain professions with faces in ways that reflect gender or ethnic discrimination. Recent works have approached the problem by rebalancing training datasets to reduce overrepresentation of dominant groups through the addition of contextualized images, for example Black female doctor or Hispanic male nurse images can help the model learn fairer associations between occupations and social categories. In contrast, we present a stealthy targeted data poisoning approach in which we add deliberately constructed training examples to the training set to minimize the biased associations the model has learned. Instead of balancing the dataset across entire categories, the training examples we add provide counter-stereotypical image–text pairs. Our goal is to reduce biased correlational models' latent representations, while still maintaining performance overall on general vision–language tasks. We use the FairFace dataset, which contains cropped face images evenly distributed across ethnicity and gender, to ensure that the model's associations between faces and professions depend solely on intrinsic facial features rather than contextual cues such as clothing, accessories, or background. By isolating the face from potentially confounding elements, we can more accurately measure the biases present in the model's latent representations and ensure that our targeted approach results in a reduction in stereotypical associations. Over-all, our work reveals how VLMs reflect societal stereotypes, thus perpetuating inequalities and making their uncritical use potentially dangerous.

Keywords: Computer Vision, Vision Language Models, Algorithmic Bias, Fairness, Data Poisoning

Syrine Enneifer is a Phd. student in Engineering in Computer Science at the Department of Computer, Control, and Management Engineering "Antonio Ruberti" at Sapienza University of Rome, under the supervision of Prof. Fabrizio Silvestri and Prof. Irene Amerini, as part of the ALCOR research group. Born and raised in Tunisia, she received her Bachelor's degree in Mathematics and Computer Science from McGill University (Canada) in 2022, and her Master's degree in Artificial Intelligence and Robotics from Sapienza University of Rome (Italy) in 2024. Her Master's thesis entitled "The perils of poison: A mathematical analysis of linear regression models under attack" received the Best Master's Thesis Award in Cybersecurity in memory of Professor Camil Demetrescu. In this work, she explored how subtle and stealthy data poisoning attacks can compromise the reliability of machine learning models in detecting fake news. Her research focuses on discrimination in data, mis/disinformation and toxic content moderation. Her academic and personal interests converge around issues of social justice, representation and the ethical use of technology.

Alessandro Turano
Dirigente scolastico, USR Calabria

Formare all'Intelligenza Artificiale nella scuola: linee guida etiche, curricolo e competenze per docenti

L'intervento propone un modello di formazione docenti per l'adozione etico-responsabile dell'Intelligenza Artificiale nella scuola, in coerenza con le recenti Linee guida ministeriali del MIM sull'IA nella didattica. Partendo dalle priorità indicate dal Ministero (trasparenza, accountability, personalizzazione, protezione dei dati, equità), l'obiettivo è delineare modalità concrete per tradurre quei principi nei curricoli, nelle pratiche didattiche e nei processi di valutazione. La proposta prevede un percorso articolato in tre moduli integrati: (1) alfabetizzazione critica all'IA – comprendere modelli, bias, limiti e opportunità per docenti non tecnici; (2) didattica con IA responsabile – progettazione di attività inclusive assistite dall'IA generativa, rubriche attive e uso consapevole degli strumenti; (3) governance scolastica dell'IA – policy locali, linee guida etiche, criteri di selezione e monitoraggio, interoperabilità e trasparenza.

Come casi d'uso didattici, si considerano applicazioni come tutoring automatico personalizzato, sistemi di feedback generativo, supporto alla scrittura riflessiva e ambienti di debate etico assistiti da IA, con attenzione agli studenti con Bisogni Educativi Speciali. Si analizzano inoltre i dilemmi centrali: delega cognitiva, bias inattesi, equità nell'accesso, opacità algoritmica. La metodologia prevede un progetto pilota con docenti ($n \approx 25$) in formazione in servizio: pre/post-test su competenze digitali e attitudini etiche, analisi qualitativa di diari riflessivi e focus group, validazione del modello con rubriche. I risultati attesi includono aumento della consapevolezza etica sull'uso dell'IA, progettazioni didattiche innovative coerenti con le linee MIM e una proposta di Framework di formazione IA-etica per la scuola italiana.

Keywords: Etica dell'IA; Formazione docenti; Curricolo; Inclusione; Cittadinanza digitale

Alessandro Turano è Dirigente scolastico, attualmente in servizio in Calabria. Laureato in Filologia classica e in Giurisprudenza all'Università Sapienza di Roma, già dottore di ricerca in Filologia e letterature greca e latina (Università G. D'Annunzio Chieti-Pescara) e in Autonomia privata curriculum realtà e radici del diritto privato europeo (Università Sapienza di Roma), è stato Cultore della materia di Diritto privato IUS-01 nella Facoltà di Giurisprudenza dell'Università Sapienza di Roma (aa.aa. 2022/24). Da settembre 2025 è distaccato presso l'Università della Calabria quale tutor organizzatore nel Corso di laurea in Scienze della Formazione Primaria. Autore di diversi contributi scientifici e saggistici nei campi del diritto, della pedagogia e della formazione, si occupa di politiche educative, innovazione curricolare e inclusione scolastica.

Alessio Vaccari
Sapienza Università di Roma

Intellectual Virtues in the Age of Artificial Intelligence: Rethinking Research Assessment

In the age of Artificial Intelligence (AI)–driven research, the assessment of scientific work should go beyond outcomes to consider the intellectual profile of researchers as well. This paper examines the role of intellectual virtues—such as open-mindedness, courage, and conscientiousness—in ensuring ethical and epistemically robust research practices. Drawing on the philosophical contributions of Sosa, Zagzebski, and Pritchard, I argue that these virtues remain essential even as AI tools, such as ChatGPT and other large

language models (LLMs), transform cognitive processes. Although AI may reduce reliance on certain internal cognitive capacities, it should not weaken intellectual virtues; on the contrary, these virtues guide researchers toward a responsible and reflective use of AI. In conclusion, I argue that, in the era of Artificial Intelligence, research assessment should place intellectual virtues at its center in order to preserve scientific integrity, foster genuine innovation, and ensure that AI acts as an ally rather than a substitute for human intellectual engagement.

Keywords: Intellectual virtues; Virtue epistemology; Responsible use of Artificial Intelligence; Research integrity

Alessio Vaccari is an Associate Professor at Sapienza University of Rome, where they teach History of Ethics and Contemporary Ethics. Their research focuses on moral philosophy and moral psychology in authors such as David Hume, Friedrich Nietzsche, and John Stuart Mill, as well as leading contemporary virtue ethicists, and on the relationship between virtue, artificial intelligence, and research assessment.

4.
DIRITTO E GOVERNANCE
LAW AND GOVERNANCE

Contributors:

1. Kasey Barr
2. Donato Cappetta
3. Wanda D'avanzo, Paola Barbara Helzel, Michele Leonetti
4. José Miguel Diéguez Rodríguez
5. Mária Fančovičová
6. Thomas Sheku Marah, Marco Marchetti
7. Andrej Mitić
8. Marianna Molinari, Ilaria Angela Amantea, Marinella Quaranta, Guido Governatori
9. Marinella Quaranta Ilaria Angela Amantea, Marianna Molinari, Guido Governatori, Marco Billi, Federico Galli, Antonino Rotolo
10. Mia Richmond
11. Giordana Truscelli
12. Mario Valori

Kasey Barr
Rubinstein Center for Constitutional Challenges, Reichman University

When Machines Advise on War: AI, Human Judgment, and Legal Responsibility under International Humanitarian Law

International Humanitarian Law (IHL) holds that lawful action in war depends on human judgment. Yet as artificial intelligence systems assist in targeting, threat assessment, and risk estimation, key questions arise: Does human judgment retain the qualities required by IHL, or is it evolving beyond of legal responsibility through technological mediation? Which aspects of judgment must remain constant to preserve accountability, and where might adaptation be both necessary and lawful? This paper investigates whether AI decision-support systems redistribute evaluative authority so extensively that judgments of proportionality, precaution, and necessity risk losing their qualitative human character. Through analysis of system functions and interaction patterns, it identifies where judgment remains strong, where it becomes vulnerable, and how these dynamics complicate regulation. The paper concludes that ensuring lawful control requires cultivating judgment integrity, a kind of behavioural and functional literacy that sustains the substance, not merely the appearance, of human judgment in technologically mediated warfare.

Keywords: AI decision-support, International Humanitarian Law, human judgment, legal responsibility, meaningful human control.

Dr. Kasey Barr is a researcher and lecturer specializing in decision-making analysis and process evaluation, with a focus on applying these methodologies to the emerging challenges of AI-assisted decision-making. Her current work at the Rubenstein Center for Constitutional Challenges (RCCC) explores the constitutional and international implications of AI systems used by powerful entities like militaries, law enforcement, and tech companies. She investigates how these technologies impact human dignity, privacy, and autonomy, and examine whether current legal frameworks are adequate to govern them. At Reichman University, she teaches courses related to decision-making in foreign policy and behavioural economics as well as academic writing for graduate students.

Donato Cappetta

Eustema SpA

Verso un'etica del Data Scientist: un approccio deontologico per l'Intelligenza Artificiale responsabile

L'evoluzione dell'Intelligenza Artificiale e della Data Science ha generato sfide etiche che richiedono un approccio sistemico e globale, superando l'etica di prossimità verso una "macro-etica". Questo intervento propone un framework deontologico per l'integrazione dell'etica nella scienza dei dati, fondato sulla Carta dei diritti fondamentali dell'Unione europea e sui recenti sviluppi normativi europei (GDPR, AI Act). Il modello proposto si articola in tre dimensioni interconnesse:

1. Etica dei Dati: governance della qualità, protezione della privacy e gestione dei metadati lungo tutto il ciclo di vita del dato
2. Etica degli Algoritmi: controllo dei bias, trasparenza decisionale e spiegabilità dei sistemi di apprendimento automatico
3. Etica delle Pratiche Professionali: responsabilità del data scientist e codici deontologici ispirati al giuramento di Ippocrate

L'approccio utilizza la metafora della piramide DIKW (Data-Information-Knowledge-Wisdom) come spazio concettuale per mappare i principi etici sui processi di knowledge discovery. Vengono analizzati casi concreti di machine-bias nei settori welfare, giustizia predittiva e marketing, evidenziando come distorsioni nei dati e negli algoritmi possano generare discriminazioni sistemiche. L'intervento propone soluzioni operative per implementare l'etica by design nei progetti di IA, bilanciando innovazione tecnologica e tutela dei diritti fondamentali. Particolare attenzione viene dedicata al ruolo del quadro normativo europeo come riferimento per lo sviluppo di sistemi IA affidabili e human-centric.

Obiettivi dell'intervento:

- Analizzare l'impatto del quadro normativo europeo sullo sviluppo di IA responsabile
- Presentare un modello operativo per l'integrazione dell'etica nella Data Science
- Proporre strumenti pratici per la prevenzione del bias algoritmico
- Stimolare il dibattito sulla necessità di una deontologia professionale per i Data scientist e tutti gli specialisti AI.

L'intervento presenta una correlazione biunivoca tra principi etici umanistici e processi matematico-scientifici tipici del data-driven, offrendo un approccio concreto per l'implementazione dell'etica come struttura integrata senza soluzione di continuità nei progetti di IA.

*

Towards an Ethics for Data Scientists: A Deontological Approach for Responsible Artificial Intelligence

The pervasive evolution of Artificial Intelligence and Big Data has generated unprecedented challenges that require overcoming traditional proximity ethics in favor of a global macro-ethics capable of governing the complexity of the infosphere. This intervention aims to outline a structured deontological framework for integrating ethics into Data Science, placing the Data Scientist at the center as the custodian of a new professional responsibility. Grounded in the Charter of Fundamental Rights of the European Union and recent regulatory developments such as the GDPR and the AI Act, the proposal establishes a bi-univocal correlation between humanistic principles and the mathematical-scientific processes typical of the data-driven approach.

The proposed operational model is articulated along three interconnected dimensions that permeate the entire data lifecycle: data ethics, understood as quality governance and privacy protection; algorithmic ethics, focused on bias control, decision-making transparency, and the explainability of machine learning models; and the ethics of professional practices. To effectively map these principles onto knowledge discovery processes, the Data-Information-Knowledge-Wisdom (DIKW) pyramid metaphor is used, transforming it into a conceptual space where ethics becomes a supporting structure rather than a mere compliance accessory. Starting from concrete cases of machine bias, the intervention highlights how distortions in data and algorithms can generate systemic discrimination, proposing as a solution also the adoption of deontological codes inspired by a 'Hippocratic Oath' for data professionals. The ultimate goal is to promote a vision of trustworthy and human-centric AI, where technological innovation proceeds hand in hand with the unwavering protection of fundamental rights.

Keywords: Data Ethics, Artificial Intelligence Act, GDPR, Bias algoritmico, Data Science deontologica, Macro-etica, Human-centric AI, Piramide DIKW

Donato Cappetta, Ingegnere Elettronico, è direttore dell'area Ricerca e Innovazione di Eustema SpA, si occupa di progetto di ricerca in ambito industriale e progetti di trasformazione digitale. Ha conseguito un master universitario di II° livello in Data Science con tesi sulle sfide etiche nell'IA. E' autore di diverse pubblicazioni scientifiche e docente in master di big data e intelligenza artificiale.

Wanda D'Avanzo, Paola Barbara Helzel and Michele Leonetti
Università della Calabria

Giustizia predittiva e intelligenza artificiale: profili etico-giuridici della disciplina vigente e riflessioni critiche

L'intelligenza artificiale assume oggi un ruolo sempre più centrale all'interno della società digitale e influenza profondamente molteplici ambiti della vita e delle attività umane. Per tali motivi, lo scopo del presente lavoro è dimostrare come l'integrazione tra esseri umani e tecnologie intelligenti, avviata dalla rivoluzione digitale, rappresenti una delle sfide più significative per la giustizia italiana, soprattutto nell'attuale fase di riforma. L'elaborato tratta il tema della giustizia predittiva, correlata da una premessa normativa di riferimento — costituita dal GDPR e dall'AI Act— che offre una tutela complessiva articolata su più livelli. Inoltre, tale interpretazione sistemica, è volta a garantire un'integrazione tra i principi giuridici vigenti e il

nuovo rapporto tra esseri umani e sistemi algoritmici. In questa prospettiva si valorizza tanto il principio di uguaglianza ex art. 3 Cost., quanto il principio di trasparenza, enunciato in Costituzione all'art. 97. Nello specifico, il presente elaborato fa riferimento agli articoli 13 e 14 della L. 132/2025, approvata il 17 settembre 2025, facenti riferimento ai sistemi di intelligenza artificiale impiegati in ambito giudiziario. Su tali enunciati, si propone al legislatore l'impiego di appositi strumenti digitali intelligenti – da prevedere nei successivi decreti attuativi- atti a salvaguardare il principio del libero convincimento del giudice. Orbene, comprendere le modalità di utilizzo dei supporti intelligenti è oggi di vitale importanza al fine di tutelare l'imparzialità del giudicato. In tal senso, in piena evoluzione non è soltanto la problematica probatoria, ma soprattutto la questione della predittività, che costituisce uno dei paradigmi evolutivi in lesione al giusto processo enunciati dall'art. 111 della Cost. In conclusione, la riflessione filosofico-giuridica sottesa dal presente lavoro intende evidenziare come l'armonizzazione del processo di integrazione tra uomo e macchina, avviato dalla rivoluzione tecnologica, rappresenti un'opportunità capace di mantenere l'essere umano al centro, in una prospettiva antropocentrica.

Keywords: e-justice; smart courts; trasparenza algoritmica e giustizia predittiva.

*

Predictive justice and artificial intelligence: ethical and legal aspects of current legislation and critical reflections.

Artificial intelligence is playing an increasingly central role within the digital society and profoundly impacts multiple areas of human life and activity. For this reason, the purpose of this paper is to demonstrate how the integration of humans and intelligent technologies, initiated by the digital revolution, represents one of the most significant challenges for the Italian justice system, especially in the current phase of reform. This paper addresses the topic of predictive justice, supported by a regulatory framework—the GDPR and the AI Act—which offers comprehensive protection across multiple levels. Furthermore, this systemic interpretation aims to ensure integration between existing legal principles and the new relationship between humans and algorithmic systems. From this perspective, both the principle of equality pursuant to Article 3 of the Constitution and the principle of transparency, set forth in Article 13 of the Constitution, are enhanced. 97. Specifically, this paper refers to Articles 13 and 14 of Law 132/2025, approved on September 17, 2025, which refer to artificial intelligence systems used in the judicial field. Based on these provisions, the legislator is proposed to use specific intelligent digital tools—to be included in subsequent implementing decrees—to safeguard the principle of free judgment. Understanding how intelligent media are used is now vitally important to safeguarding the impartiality of judgments. In this sense, not only is the evidentiary issue in full evolution, but above all the question of predictability, which constitutes one of the evolutionary paradigms undermining the fair trial enshrined in Article 132 of the Italian Civil Code. 111 of the Constitution. In conclusion, the philosophical and legal reflection underlying this work aims to highlight how the harmonization of the integration process between man and machine, initiated by the technological revolution, represents an opportunity capable of keeping the human being at the center, in an anthropocentric perspective.

Keywords: e-justice; smart courts; AI ACT; GDPR

Wanda D'Avanzo è avvocato e assegnista di ricerca in Filosofiadel diritto. Esperta di Informatica giuridica e diritto dell'informatica, ha insegnato in diversi atenei, tra cui l'Università di Roma Unitelma Sapienza e l'Università degli Studi Magna Graecia di Catanzaro. Attualmente è docente di Logica del diritto e informatica giuridica presso l'Università della Calabria. Autrice di numerosi saggi scientifici su varie riviste nazionali

e internazionali, ha da ultimo pubblicato per Rubbettino editore: "E-government e società del controllo. Il lato oscuro del mondo digitale" (2023) e "Il Metaverso. Le nuove frontiere della tecnologia tra etica e diritto" (2025).

Paola Barbara Helzel è professore Associato di Filosofia del diritto nel CdS in Giurisprudenza dell'Università della Calabria dove insegna Biogiuridica, Logica e argomentazione del diritto e Filosofia del Diritto. Autrice e coautrice di pubblicazioni a livello nazionale e internazionale. Tra le sue pubblicazioni: Il diritto ad avere diritti. Per una teoria normativa della cittadinanza (2005), Per una teoria generale del dovere (2016). In qualità di curatrice ha pubblicato Dalla Galassia al sistema. La ricerca dell'ordine (2018); Quale Bioetica per il terzo millennio? (2021); in co-curatela con F. Rubio Sanchez, Sport e minori tra etica e diritto (2024), I. C. Pasca, Diritto ed economia tra tradizione e innovazione (2024).

Michele Leonetti è dottorando di ricerca presso l'Università della Calabria, dove svolge attività di studio e approfondimento nell'ambito dell'Informatica giuridica e dell'intelligenza artificiale applicata al diritto. È cultore della materia per gli insegnamenti di Filosofia del diritto e di Informatica giuridica presso la medesima Università. Ha conseguito una fellowship presso la Fondazione Vaticana "Centesimus Annus – Pro Pontefice". Attualmente svolge funzioni di supporto legislativo presso il Senato della Repubblica, nell'ambito dell'VIII Commissione permanente. È membro delle seguenti società scientifiche: SIpEIA (Società Italiana per l'Etica nell'Intelligenza Artificiale), AIxIA (Associazione Italiana per l'Intelligenza Artificiale) e ANDIG (Associazione Nazionale Docenti di Informatica Giuridica e Diritto dell'Informatica).

José Miguel Diéguez Rodríguez
University of Murcia

Gaining time to do the right thing or simply missing the point? The inclusion of stress and burnout as "emotions" in the Guidelines on AI Prohibited Practices

The Guidelines on Prohibited Artificial Intelligence Practices issued by the European Commission interpret Article 5 of the AI Act as extending the general ban on emotion-recognition systems to those detecting "stress or burnout". Although grounded in a legitimate effort to prevent invasive or discriminatory uses of AI, this approach reflects a conceptual misunderstanding of stress and related conditions and risks disregarding the practical value that their careful measurement could have for improving occupational safety and health in the workplace. By treating stress as a mere emotion, the Guidelines overlook its well-established association with serious medical conditions, particularly cardiovascular disease, which reveals its complex interaction between psychological and physiological factors. In the same vein, the International Classification of Diseases defines burnout as an occupational syndrome rather than an emotional state. The Guidelines envisage limited exceptions, such as assistive systems supporting autistic individuals. Yet for stress or anxiety monitoring, the only cases acceptable under their reasoning concern the prevention of accidents, such as when operating heavy machinery or handling hazardous substances, thereby excluding preventive or diagnostic applications. The paradox is evident: technology may assist those already diagnosed but cannot help identify or protect undiagnosed workers whose performance or safety may suffer precisely because no adaptation was made. Other benefits could include supporting ergonomic assessments or providing evidence for workers and their representatives that the pace or organisation of work harms health through objective stress indicators.

However, it is undeniable that these techniques could be misused against workers' rights, with serious consequences—a risk that cannot be ignored. Yet, with proper safeguards, their ethical use could strengthen occupational health and safety. The question, then, is whether the Commission's stance is a temporary precaution allowing time for more nuanced regulation, or a lasting prohibition that misses the chance to better protect workers.

Key words: AI Act; emotion recognition; occupational safety and health; stress monitoring; human-centred governance

José Miguel Diéguez Rodríguez is a PhD candidate at the University of Vigo. His research focuses on the impact of neurotechnologies on fundamental rights, particularly in the field of labour and social rights. He is currently collaborating on a project at the Autonomous University of Madrid concerning the implementation of the principles on neuro-rights, as set out in the Spanish Charter of Digital Rights, into specific legislative proposals, where he is responsible for the chapter on labour law.

Mária Fančovičová

Department of Interna-onal Law and European Law, Faculty of Law, Palacký University

Artificial Intelligence And The Concept Of The Moral Imperative

In the context of ongoing discussions on the AI revolu5on, not only in military affairs, where part of the international community calls for a complete ban on fully autonomous weapons systems (AWS), the concept of the moral imperative for minimally-just autonomy using AI (MinAI) argues that imposing a total ban on the development of AI in weapon systems is morally unjustifiable, as AI has the potential to reduce human suffering. The paper therefore examines in detail the concept of MinAI systems, which are designed to prevent aGacks on persons and objects protected under international humanitarian law. In contrast, maximally-just autonomy using artificial intelligence (MaxAI) systems aims to make autonomous ethical decisions and perform "morally right" actions, whereas MinAI represents a simpler approach designed solely to prevent ethically and legally impermissible actions. In conclusion, although current AI is still unable to make responsible moral decisions comparable to those of humans, MinAI can nevertheless be used e.g. to idenfy legally protected emblems (e.g., the Red Cross) and to abort aGempts to aGack medical facilities, thereby reducing harm to civilians.

Keywords: artificial intelligence; the concept of the moral imperative; ethics; decision-making

Mária is a doctoral student at the Department of International Law and European Law, Faculty of Law, Palacký University, Czech Republic. Her research focuses on autonomous weapons systems (AWS) within international law, with particular emphasis on international humanitarian law (IHL). She examines existing legal frameworks governing AWS and explores pathways for their future regulation. Her work also addresses the ethical and security challenges raised by the integration of artificial intelligence into military decision-making, while engaging more broadly with AI from a legal perspective. In the context of ongoing debates surrounding the AI revolution, including calls from parts of the international community for a comprehensive ban on fully autonomous weapons, Mária analyzes in one of her research projects the concept of a moral imperative for minimally just autonomy using artificial intelligence (MinAI). Her research exam-

ines the concept of MinAI systems, which are designed to prevent attacks on persons and objects protected under IHL. These systems stand in contrast to maximally just autonomy (MaxAI), which seeks to enable autonomous ethical decision-making and morally "right" actions. In conclusion, although current AI is still unable to make responsible moral decisions comparable to those of humans, MinAI can nevertheless be used, for example, to identify legally protected emblems (e.g. the Red Cross) and to abort attempts to attack medical facilities, thereby reducing harm to civilians.

Thomas Sheku Marah and Marco Marchetti
Department of Law, Nusa Putra University

AI, Sovereignty, and Global Governance: Ethical and Legal Pathways for Human-Centered Regulation

Artificial Intelligence revolutionizes governance alongside law and international relations but regulatory systems exist in fragmented and uneven patterns. National frameworks including the European Union's AI Act demonstrate citizen protection attempts which reveal fundamental conflicts between state sovereignty and necessary international responsibility coordination. The research evaluates ethical along with legal obstacles that emerge during human-oriented AI governance. The paper demonstrates that independent and regional policy frameworks fail to address cross-border dangers which include algorithmic prejudices and false information distribution and cybersecurity risks. Global cooperation stands essential for establishing inclusive standards which protect human dignity while respecting cultural diversity.

The research uses comparative analysis of international initiatives including UNESCO's Recommendation on the Ethics of Artificial Intelligence together with African Union guidelines and Asian regulatory experiments to show normative approach similarities and differences. These developments get placed into wider discussions about sovereignty and accountability and inclusivity as ethical AI governance needs to move beyond compliance-based legal frameworks. A collaborative values-based approach that combines legal principles with diplomatic relations and ethical considerations provides the strongest foundation for building trust and accountability and protecting technological innovation resilience. The paper presents a global governance framework which connects national priorities to worldwide obligations. When regulation integrates ethical standards and people from different regions start talking about AI governance it will promote technological development while maintaining fundamental human rights protection. To achieve sustainable and equitable innovation during the twenty-first century a human-centered approach serves as an essential element.

Keywords: AI Regulation; Sovereignty; Global Governance; Ethics; International Law

Thomas Sheku Marah is an undergraduate student of International Law at Nusa Putra University, Indonesia, originally from Sierra Leone. He is the founder of the Salone Future Leaders Foundation. His work focuses on youth empowerment, global governance, and the intersection of law, policy, and technology. He has represented Sierra Leone at international cultural and academic programs, authored articles on governance and youth participation, and is committed to advancing ethical, inclusive, and human-centered approaches to global challenges.

Marco Marchetti is a policy advisor and university lecturer with expertise in comparative politics, international governance, and institutional relations. His career spans Europe, Africa, and Southeast Asia, where

he has worked to connect academia, public institutions, and the private sector. He is committed to shaping and supporting institutions and policy platforms that drive high-impact, evidence-based public decisions. Inspired by the model of the Tony Blair Institute for Global Change, his approach combines academic depth with practical policy implementation in global governance and institutional reform.

Andrej Mitić
Department of Pirot, Academy of Technical and Pedagogical Vocational Studies

Law and Artificial Intelligence Between Normative Convergence and Geopolitical Divergence: The Case of Serbia in the Context of the EU AI Act

The rapid development of Artificial Intelligence (AI) presents one of the most complex challenges for contemporary legal and policy frameworks. While the European Union has established itself as a global pioneer in normative AI governance through the adoption of the AI Act, candidate countries such as the Republic of Serbia face the dual challenge of alignment with EU standards and preservation of their institutional and geopolitical autonomy. Situated at the intersection of the EU, USA, and China's technological and political influences, Serbia's approach to AI regulation cannot be viewed merely as a process of legal harmonization but as a broader strategic positioning within competing models of digital sovereignty.

This paper examines the evolving Serbian legal and strategic framework for AI, analyzing its trajectory in the context of the EU AI Act and related policy instruments. The research employs a three-layered methodology: a normative analysis of EU documents; a review of Serbian strategies and policy papers, including the National AI Strategy 2025–2030; and a comparative assessment of legal convergence between Serbia and the EU. Additionally, a critical-theoretical approach is applied to explore the ethical foundations necessary for regulating algorithmic decision-making and ensuring accountability. Preliminary findings suggest that Serbia's regulatory evolution remains primarily strategic and declarative, reflecting aspirations for European integration, while balancing technological cooperation with China. The key question, therefore, is not only how Serbia can transpose EU legal mechanisms but how it can translate them into an operational, context-sensitive framework capable of addressing local institutional limitations and societal expectations.

The contribution of this paper lies in proposing a contextualized model of responsible AI governance for Serbia—one that integrates European normative principles with national specificities and recognizes the geopolitical complexity of technological regulation in Southeast Europe.

Keywords: AI regulation, EU AI ACT, Serbian legal framework, Geopolitics of technology, Responsible AI governance

Andrej Mitić graduated in Philosophy from the University of Belgrade, and holds a PhD in Legal Sciences from the Faculty of Law, University of Niš, Serbia. He currently teaches Law & Artificial Intelligence at the Academy of Applied Technical and Preschool Studies, University of Niš. His academic experience includes a Europe-Next-to-Europe Fellowship at the New Europe College – Institute for Advanced Study in Bucharest, as well as a research stay at the European University Institute (EUI) in Florence. His research interests include legal and political theory, political philosophy, and law, ethics, and artificial intelligence. He is the author of several scholarly publications and has participated in numerous international scientific conferences in the European Union and the Republic of Serbia.

Marianna Molinari
LaST-JD, Legal Studies Dept., Alma Mater Studiorum Università di Bologna
Computer Science Dept.,
Università degli studi di Torino
PREC Dept., Vrije Universiteit Brussel

Ilaria Angela Amantea
Computer Science Dept.,
Università degli studi di Torino

Marinella Quaranta
LaST-JD, Legal Studies Dept., Alma Mater Studiorum Università di Bologna
Computer Science Dept.,
Università degli studi di Torino
PREC Dept., Vrije Universiteit Brussel

Guido Governatori
Central Queensland University

Employing LLMs to Extract Principles of Law for Predicting Outcomes: An Analysis of Italian Judgments on LGBTQIA+ Rights

Principles of Law (PoLs) are essential legal concepts that adapt to societal needs, especially in areas lacking specific legislation, such as LGBTQIA+ rights. Judges often create these principles when existing laws do not adequately address new legal challenges. By establishing PoLs through their rulings, they form a framework that ensures consistency in deciding across similar cases. This process helps maintain legal certainty, guarantees fair treatment, and provides guidance on how future cases might be handled. However, PoLs are not always clearly defined in rulings, and their expression can vary, making it difficult to identify them without detailed analysis. As judgments come in different formats, understanding the core PoLs requires careful interpretation. On the other hand, timely identification of these principles is crucial for facilitating swift decisions and helping citizens adjust their conduct in accordance with evolving legal standards. This study builds on previous work by incorporating several Large Language Models (LLMs) and comparing their performance against a baseline using Regular Expressions (RegEx).

The methodology, involving a Python script generated by ChatGPT, ensures that the process is reproducible and verifiable. The research evaluates the effectiveness of LLMs, including ChatGPT, Claude, and DeepSeek, in extracting PoLs from Italian court rulings related to LGBTQIA+ rights. Results were compared with extractions obtained using RegEx. While LLMs showed potential, they fell short in both quantity and accuracy, underscoring the importance of expert oversight. The findings highlight that while LLMs can summarize judgments, they remain prone to biases, inaccuracies, and confusion, especially when distinguishing between PoLs and irrelevant legal content.

Keywords. AI and Law, Principles of Law, LLMs, RegEx, Extraction.

Marianna Molinari is a joint Doctoral Researcher in Law, Science and Technology at the University of Bologna, University of Turin and Vrije Universiteit Brussel. She obtained her graduation in Law from the University of Siena, with honors. In particular, she focused on new frontiers of Family Law, getting an LSS in Comparative Family Law at the University Carlos III of Madrid, jointly with the Seattle Law School and an LLM in Family Law at LUISS University of Rome. Following her tenure at the University of Turin as both a research scholar and grant-holder, she transitioned to her current role at the University of Bologna, University of Turin and Vrije Universiteit Brussel. Her ongoing research explores the emerging field of predictive justice, delving into its potential application to cases concerning personal rights. In addition to her research, she brings practical experience as licensed to practice law by the Italian Bar and former auditor at an Italian Court of First Instance.

Marinella Quaranta is a Ph.D. candidate in the "Law, Science and Technology" (LAST JD) program at the University of Bologna, jointly affiliated with the University of Turin and Vrije Universiteit of Brussel. She holds a Master's Degree in European Legal Studies from the University of Turin. Marinella's academic interests lie primarily in Legal Informatics, particularly the implications of artificial intelligence (AI) in the medical field. Her master's thesis examined the development of European Regulation on AI systems and its interaction with European medical frameworks, in collaboration with "Città della Salute" in Turin. Marinella currently works on AI compliance with national and international legislative frameworks in Europe. Her research focuses on AI compliance within national and international legislative frameworks, with particular emphasis on the implementation of the AI Act. The central objective of her doctoral project is to design a general compliance structure that supports both general and healthcare AI providers in aligning with EU and national requirements. She is involved in the AI Factory consultancy project, "ITA4LIA AI Factory-Italy for Artificial Intelligence" within its legal unit "Legal triage". The project provides AI stakeholders with guidance throughout the compliance process, from pre-market assessment to post-market monitoring.

Marianna Molinari
LaST-JD, Legal Studies Dept., Alma Mater Studiorum Università di Bologna
Computer Science Dept.,
Università degli studi di Torino
PREC Dept., Vrije Universiteit Brussel

Ilaria Angela Amantea
Computer Science Dept.,
Università degli studi di Torino

Marinella Quaranta
LaST-JD, Legal Studies Dept.,
Alma Mater Studiorum Università di Bologna
Computer Science Dept.,
Università degli studi di Torino
PREC Dept., Vrije Universiteit Brussel

Guido Governatori
Central Queensland University

Marco Billi
LaST-JD, Legal Studies Dept.,
Alma Mater Studiorum Università di Bologna

Federico Galli
LaST-JD, Legal Studies Dept.,
Alma Mater Studiorum Università di Bologna

Antonino Rotolo
LaST-JD, Legal Studies Dept.,
Alma Mater Studiorum Università di Bologna

Designing A Compliance Tool in Light of the EU AI Act

The AI Act formally entered into force on 1 August 2024, establishing a two-year transitional period before the Regulation becomes fully applicable on 2 August 2026. However, certain obligations will take effect at later stages. Specifically, the compliance deadline for high-risk AI systems has been extended to August 2027, while full compliance across the EU technology sector is expected by 2030. This work aims to provide a Handbook that proposes a structured and conceptual framework for an ideal compliance tool, focusing specifically on the obligations applicable to high-risk AI systems from the provider's perspective. To this end, we identified and analyzed the relevant provisions of the AI Act, organizing them according to the distinct phases of an AI system's lifecycle. For each phase, we conceptualized a set of compliance tasks

designed to address the specific regulatory requirements. Subsequently, we defined corresponding verification tasks that assess whether each compliance activity has been properly fulfilled. The outcome of this work is a comprehensive Handbook that systematically presents the obligations imposed by the AI Act across the three main phases of an AI system's lifecycle, project design, development and deployment, and monitoring. Moreover, it reverses the perspective: rather than merely listing obligations, it delineates how an ideal compliance tool should operate to verify adherence to these requirements. For each obligation, the Handbook details the specific verification and compliance tasks associated with each phase, thereby offering a methodical and operational reference for ensuring conformity with the AI Act.

Keywords. AI and Law, EU AI Act, AI Compliance Methodology, AI Act Compliance.

Marianna Molinari is a joint Doctoral Researcher in Law, Science and Technology at the University of Bolo-gna, University of Turin and Vrije Universiteit Brussel. She obtained her graduation in Law from the University of Siena, with honors. In particular, she focused on new frontiers of Family Law, getting an LSS in Comparative Family Law at the University Carlos III of Madrid, jointly with the Seattle Law School and an LLM in Family Law at LUISS University of Rome. Following her tenure at the University of Turin as both a research scholar and grant-holder, she transitioned to her current role at the University of Bologna, University of Turin and Vrije Uni-versiteit Brussel. Her ongoing research explores the emerging field of predictive justice, delving into its poten-tial application to cases concerning personal rights. In addition to her research, she brings practical experience as licensed to practice law by the Italian Bar and former auditor at an Italian Court of First Instance.

Marinella Quaranta is a Ph.D. candidate in the "Law, Science and Technology" (LAST JD) program at the University of Bologna, jointly affiliated with the University of Turin and Vrije Universiteit of Brussel. She holds a Master's Degree in European Legal Studies from the University of Turin. Marinella's academic interests lie primarily in Legal Informatics, particularly the implications of artificial intelligence (AI) in the medical field. Her master's thesis examined the development of European Regulation on AI systems and its interaction with Eu-ropean medical frameworks, in collaboration with "Città della Salute" in Turin. Marinella currently works on AI compliance with national and international legislative frameworks in Europe. Her research focuses on AI com-pliance within national and international legislative frameworks, with particular emphasis on the implementa-tion of the AI Act. The central objective of her doctoral project is to design a general compliance structure that supports both general and healthcare AI providers in aligning with EU and national requirements. She is in-volved in the AI Factory consultancy project, "ITA4LIA AI Factory-Italy for Artificial Intelligence" within its legal unit "Legal triage". The project provides AI stakeholders with guidance throughout the compliance process, from pre-market assessment to post-market monitoring.

Mia Richmond
University of Cambridge

The Legal Status of Psychographic Profiling

Increasingly sophisticated psychographic profiling methods provide platforms with the power to predict and influence human behaviour. The scope is totalizing: what we buy, whom we vote for, what we watch, whom we love, what makes us afraid. The European Union's General Data Protection Regulation

(GDPR) seeks to equip individuals with rights intended to restore control and contestability in the face of such data-driven influence. The rights it entails are more than procedural entitlements; they embody a particular vision of the relationship between individuals and information systems.

Yet in this paper, I ask: does the way that we regulate psychographic profiles align with their technical reality? I argue that it does not, focusing on GDPR-style rights. This misalignment is structural rather than a failure of enforcement, explanation, or institutional capacity. While the law implicitly treats psychographic profiles as stable personal-data objects, in practice psychographic profiling reduces data subjects to fluid coordinates in high-dimensional latent space – representations that are relational, probabilistic, distributed, and dynamically instantiated. This means that GDPR-style individual rights, such as access, rectification, and erasure, are often rendered conceptually inapplicable or practically unexercisable.

I situate this work within existing scholarship and then identify four assumptions about personal data embedded in GDPR-style rights: that profiles are individuated, truth-apt, locatable, and stable. I contrast these assumptions with the technical reality of psychographic profiling, demonstrating why core data subject rights become conceptually inapplicable upon closer examination. Finally, I consider the regulatory implications of this mismatch for data protection law, and for legal approaches to governing profiling and behavioural influence more generally.

Keywords: data rights, cognitive sovereignty, autonomy

Mia Richmond is currently completing her MPhil in the Ethics of AI, Data, and Algorithms at the University of Cambridge, supported by a scholarship from the Fondazione Randstad- AI & Humanities. Her research focuses on data protection frameworks and the protection of autonomy and cognitive sovereignty in AI governance. She also works as a research assistant at the Uehiro Oxford Institute and has held positions with the Brennan Center for Justice at NYU Law and the American Philosophical Association. Mia previously studied philosophy and psychology at Columbia University.

Giordana Truscelli
Università degli Studi di Teramo

Ius e Lex: quando l'algoritmo detta legge

L'integrazione dell'intelligenza artificiale nell'amministrazione pubblica costituisce una sfida importante sia per la filosofia del diritto sia per il diritto amministrativo, in quanto pone spunti di riflessione che investono la natura stessa della norma giuridica e la dimensione antropologica del diritto. La tensione fondamentale che si riscontra è nel rapporto tra lex e ius: mentre l'IA, in via di principio, potrebbe eccellere nell'applicazione meccanica della norma positiva codificata, a causa del suo funzionamento basato su pattern ricorrenti e calcolo computazionale, appare inidonea a cogliere lo ius, inteso come giusto concreto che emerge dal caso particolare e che richiede quella prudentia iuris, identificata come elemento essenziale di un'esperienza giuridica autentica. La distinzione tra ius e lex attraversa millenni di dibattito dottrinale, che distingue lo ius, ancorato ad una dimensione di giustizia radicata nell'etica ed insito nella natura delle cose, dalla lex, intesa al contrario come norma codificata, la lettera della legge, applicata e sanzionata.

L'intento di questo contributo è quello di mettere in luce i diversi profili problematici per contribuire ad un autentico dibattito su di essi nell'era della società digitale. L'automazione decisionale ed il problema della trasparenza algoritmica assumono quindi, una connotazione più profonda, costituendo forse una sorta di alienazione giuridica che trasforma il soggetto di diritti in oggetto di elaborazione algoritmica. La

sfida che attende la nostra società consiste nel garantire che l'IA non si limiti ad applicare rigidamente la lex rischiando di perpetrare ingiustizie e discriminazioni in nome di una oggettività algoritmica che vorrebbe realizzare l'antico bisogno dell'essere umano di raggiungere "la certezza del diritto". Il problema centrale, dunque, è quello di interrogarsi su cosa accade al diritto quando la decisione viene delegata alle "macchine" e se il diritto è un'arte al servizio dell'umano.

Keywords: decisione pubblica, giustizia, intelligenza artificiale.

*

Ius and Lex: when the algorithm dictates the law

The integration of artificial intelligence in public administration constitutes a significant challenge for the philosophy of law and administrative law, as it gives rise to issues that impact the very nature of legal norms and the anthropological dimension of law. The fundamental tension that exists pertains to the relationship between lex and ius: while AI, in principle, could excel in the mechanical application of the codified positive norm, because of its operation based on recurring patterns and computational calculation, it appears to be unsuitable for grasping the ius, understood as the concrete right emerging from the particular case and requiring 83ocused8383z iuris, identified as an essential element of authentic legal experience. The distinction between ius and lex has been a subject of extensive debate within the field of jurisprudence over the course of millennia. Ius, understood as an abstract concept of justice embedded within ethical principles and inherent in the very nature of things, differs from the lex, which is regarded as a codified norm, the letter of the law, that is applied and sanctioned. The purpose of this contribution is to draw attention to the various problematic profiles in order to contribute to a more informed discussion. A discourse on the subject in the contemporary digital era. The implications of automation in decision-making and the issue of algorithmic transparency thus assume a heightened significance, potentially engendering a form of legal alienation that metamorphoses the subject of rights into the object of algorithmic processing. The challenge confronting our society is to ensure that AI does not become confined in its application of the law, thereby risking the perpetuation of injustice and discrimination in the pursuit of algorithmic objectivity. This pursuit is predicated on the fulfilment of an age-old human need to achieve 'legal certainty'. The fundamental issue, therefore, is to interrogate the implications of delegating legal decision-making to 'machines' and to explore the question of whether law can be considered an art that serves humankind.

Keywords: public decision-making, justice, artificial intelligence.

Licensed lawyer, Graduated in Law from LUISS Guido Carli and the Political Science University of Teramo. Specialised in Administrative Law and Science of Administration. She is currently enrolled as a PhD student in European Studies for Innovation at the University of Teramo, where she also serves as a phd student in the Philosophy of Law and Administrative Law. The individual has been awarded the first prize at the PhD Annual Meeting and is a member of the working group. Her research, as a member of the CNR 'AI and Urban Intelligence' research group, focuses on the intersections between artificial intelligence, the philosophy of law and administrative law. She has participated as a speaker at national and international congresses and has been published in scientific journals in band A and collective volumes, both nationally and internationally.

Mario Valori
Università di Pisa

Apertura vincolata: AI Act, licenze etiche e sostenibilità dell'open source tra Europa e Sud globale

L'AI Act riconosce un ruolo privilegiato e notevoli semplificazioni ai prodotti "free & open-source", ma subordina tali benefici a condizioni che riplasmano l'identità dell'open source nello spazio europeo. Questa proposta analizza come le esenzioni promosse e il vincolo di assoluta gratuità prevista per l'open source interagiscano con pratiche comunitarie e di mercato europee; in aggiunta si analizzerà il fenomeno, a oggi emergente ma comunque già significativo, delle licenze 'etiche' (es. RAIL, Hippocratic) che introducono limiti d'uso responsabile per i propri prodotti.

Leggendo normativa e licenze come artefatti tecno-politici, la domanda è duplice: 1. In che modo l'inquadramento giuridico europeo incentiva o scoraggia forme di apertura economicamente sostenibili, specie per attori piccoli e periferici? 2. Fino a che punto le licenze etiche possono operare come strumenti di governance emergente senza rinnegare i valori FOSS di libertà d'uso che hanno sempre contraddistinto questo ecosistema e senza diventare strumenti di imposizione di valori coloniali predeterminati?

Lo studio si avvale di analisi documentale (testi normativi, linee guida, testo delle licenze), di comparazione fra casi reali e pratici, di uno studio dei pareri manifestati dalla comunità open-source attiva online e di una lettura socio-giuridica delle disgiunzioni tra definizioni istituzionali e pratiche effettive di adozione. Il risultato presenta una prima mappatura dei punti di attrito fra 'apertura' formale e sostenibilità, accompagnata da una valutazione dell'efficacia e della legittimità delle clausole etiche nel prevenire usi qualificati come nocivi o dual-use ambigui, tentando anche di individuare tentativi di 'openwashing' nei quali l'apertura diviene mero marchio conformativo. Una particolare attenzione viene riservata a come tali scelte possano influenza il sud globale, sia come importatori di prodotti open-source europei che come esportatori verso l'Europa di software originale che, però, pur essendo open- source secondo i criteri FOSS, potrebbe non esserlo secondo la normativa europea.

Keywords: Open source; AI Act; Licenze etiche; Sostenibilità economica; Pluralismo normativo

Mario Valori ha conseguito la Laurea magistrale in giurisprudenza (2008) e in scienze amministrative (2012); fin dalla laurea ha lavorato nel campo della trasformazione digitale, del diritto delle nuove tecnologie e della proprietà industriale, con un focus sulla normativa europea e transnazionale. Con l'avvento dell'AI e i forti cambiamenti che ha portato, si è reso necessario un nuovo periodo di formazione, con una specializzazione in etica delle tecnologie digitali e intelligenza artificiale (2024) e una tesi sull'addestramento etico e di qualità dei dataset per la NMT. Da allora, il lavoro si è concentrato nella selezione dei dati per l'addestramento delle AI generative, nella creazione di strutture etiche d'uso, nelle funzioni di Data Protection Officer e di redteamer per i prodotti AI generativi. Strenuo fautore dell'open-source e dell'open-access, negli anni ha partecipato a conferenze internazionali sia tecniche che accademiche; fra le più recenti, un'analisi delle procedure di sicurezza nella struttura IT europea durante la pandemia (DEEPSEC, Austria), uno studio sulla percezione fra il pubblico delle digital humanities (AIUCD, Italia) e un'analisi dei bias di genere nella Neural Machine Translation (University of the Aegean, Grecia).

5.

RESPONSABILITÀ E CURA
ACCOUNTABILITY AND CARE

Contributors:

1. Jennifer Ang
2. Martina Baltuzzi
3. Alessandro Brusadelli
4. Eleonora Catena
5. Yalena Chothia
6. Julia Egenhoff
7. Emily Elstub
8. Petr Jošt
9. Ryan SangBaek Kim

Jennifer Ang

Singapore University of Social Sciences

What is it about Care that Carebot Changes?

This paper considers the question of care and technology. Carebots are designed to take care of the care receiver by assisting with a range of caregiving needs. What carebots are not expected to be capable of, at least for now, is to care about a care receiver, that is, to have a subjective state of concern, or a humanistic care 85ocused85 to care for a care receiver by looking out for their interests. This perspective thus assumes that carebots are tools that assist with care tasks, but it overlooks the fact that in the care space, tasks are not only part and parcel of care practice but also demonstration of the human relationality of care. What are the affordances and constraints of carebots? How will carebots through taking over care tasks alter how we care about and care for care receivers?

Keywords: care; carebots; care practice; moral deskilling

Jennifer Ang is Associate Professor of Philosophy at the Singapore University of Social Sciences. She is the author of Justified Forgiveness: A Moral Case against Neglect, Silence and Forgetfulness (Bloomsbury 2025) and Sartre and the Moral Limits of War and Terrorism (Routledge 2010, 2014). Currently, she is the lead-PI of 2 funded projects on AI – "Ethics of Autonomy and Care in AI Decision-making" with the multidisciplinary Singapore-France research programme Decision-making of Urban Critical Systems (DesCartes) at CNRS@CREATE, and "Technologization of Singapore(ans)" with the Southeast Asia Neighbourhood Network (SEANNET). Of relevance to this conference submission is the journal publication "The Problem of the Responsibility of Nobody," Divus Thomas. 126 (2), 176-199.

Martina Baltuzzi

Università degli studi di Torino | Alma Mater Studiorum Università di Bologna

Digital Twins in Medicine: Navigating Technological Autonomy and the Ethics of Care

Digital twins (DTs) are virtual replicas of tangible products or processes which, utilising data streams to generate digital representations of the physical counterpart, can simulate and mirror real-world alterations. Since the early 2000s, when Michael Grieves first introduced them at the University of Michigan, they have experienced increasing popularity, holding the promise of revolutionizing personalized medicine by enhancing diagnosis, treatment, and prevention in various health areas. Yet, their implementation raises questions about the relationship between technological innovation and care. On the one hand, Digital Twins could "bring data regarding the patient closer to the patient" serving as empowering tools and allowing individuals to exercise a greater degree of autonomy over healthcare decisions. On the other hand, as Zullo has observed, the notion of patient autonomy has, over the past decade, been reconsidered through the lens of the ethics of care. This perspective emphasizes that moral and professional responsibility extends beyond respecting autonomy to include providing supportive, compassionate, and communicative care. Mortari's philosophy of care further highlights the universality of human interdependence and the moral significance of shared emotions and harmonious relations. Within this context, the ethical discourse surrounding Digital Twins must ensure that technological empowerment does not erode the relational and affective dimensions fundamental to care. Rather, the integration of DTs in medicine should reinforce a model of healthcare that harmonizes technological innovation with enduring principles of empathy, dignity, and relational responsibility. The purpose of this research is to analyse the current discourse surrounding DTs in medicine, underlining the importance of the ethics of care perspective and addressing the risks of depersonalising healthcare through technology.

Keywords: Digital twins, medicine, ethics of care

Martina Baltuzzi is a PhD candidate in Law, Science and Technology Joint Doctorate, 40th cycle, at the Universities of Bologna and Turin. Her research project, conducted under the supervision of Professors Ugo Pagallo and Massimo Durante, is entitled "Digital Twins in Healthcare: Norms, Epistemology and Ethics in Europe". She previously held a research fellowship at the Law Department of the University of Turin, focusing on the ethical and legal challenges of autonomous mobility and algorithmic discrimination. She graduated in European Legal Studies at the University of Turin, completing part of her master's thesis at the Technical University of Munich.

Alessandro Brusadelli

Dipartimento di Scienze Cliniche e Sperimentali, Università degli Studi di Brescia

Dati sintetici e identità computazionale: ripensare l'evidenza e la governance dell'IA in sanità.

L'introduzione dei dati sintetici nella ricerca clinica segna un punto di svolta nel modo in cui la medicina produce conoscenza. Finora, l'uomo ha fondato la conoscenza clinica sull'osservazione e sulla raccolta empirica di dati: l'IA, invece, inaugura una conoscenza che non osserva, ma genera. Con i dati sintetici, la realtà

non è più semplicemente rilevata, ma modellata, e la verità empirica cede il passo a una verità computazionale, costruita attraverso algoritmi. Da questa trasformazione epistemica emerge il concetto di identità computazionale, intesa non come nuova forma dell'essere, ma come nuova struttura del conoscere: l'individuo non si limita più a essere rappresentato nei dati, ma viene ricostruito attraverso processi di generazione artificiale che producono pazienti possibili e scenari di evidenza simulata.

In questa prospettiva, il dato sintetico non è una mera tecnica di anonimizzazione, ma un atto di produzione di conoscenza, che interpella il diritto dei dati personali nei suoi fondamenti: identità, referenzialità, consenso, responsabilità. Il contributo propone di leggere tali trasformazioni alla luce del Regolamento (UE) 2021/2282 (HTA) e dell'art. 8 del DDL IA, mostrando come la governance etica dell'IA sanitaria debba fondarsi su un equilibrio tra libertà della ricerca, tutela della persona e interesse pubblico alla conoscenza. L'obiettivo è delineare un quadro in cui la protezione dei dati non sia il limite, ma la condizione generativa della verità scientifica, restituendo al diritto il compito di ordinare la nuova forma della conoscenza nell'età algoritmica.

*

The introduction of synthetic data into clinical research marks a turning point in the way medicine produces knowledge. Until now, clinical knowledge has been grounded in observation and the empirical collection of data; artificial intelligence, by contrast, inaugurates a form of knowledge that does not merely observe but generates. Through synthetic data, reality is no longer simply recorded but modeled, and empirical truth gives way to a computational truth constructed through algorithms.

From this epistemic transformation emerges the concept of computational identity, understood not as a new mode of being, but as a new structure of knowing. The individual is no longer merely represented within data but is reconstructed through processes of artificial generation that produce possible patients and simulated scenarios of evidence.

From this perspective, synthetic data should not be regarded as a mere anonymization technique, but rather as an act of knowledge production that challenges the foundational categories of personal data law: identity, referentiality, consent, and responsibility. The contribution proposes to interpret these transformations in light of Regulation (EU) 2021/2282 (HTA) and Article 8 of Law No. 132/2025, showing how the ethical governance of AI in healthcare must be grounded in a balance between the freedom of scientific research, the protection of the person, and the public interest in knowledge production.

The aim is to outline a framework in which data protection does not constitute a limit, but rather the generative condition of scientific truth, restoring to law the task of ordering the new form of knowledge in the algorithmic age.

Keywords: Dati sintetici; sperimentazione clinica; intelligenza artificiale; identità; protezione dei dati personali

Alessandro Brusadelli è dottorando in Intelligenza Artificiale in Medicina e Innovazione nella Ricerca Clinica e Metodologica presso l'Università degli Studi di Brescia. Ha coordinato un gruppo di ricerca multidisciplinare sull'impiego dell'uso dei dati sintetici in ambito sanitario, contribuendo alla pubblicazione di "Ripensare l'evidenza: approccio multidisciplinare all'uso dei dati sintetici in ambito sanitario" su BioLaw Journal. I suoi interessi riguardano le implicazioni giuridiche e filosofiche dell'IA in medicina.

Eleonora Catena
Centre for Philosophy and AI Research {PAIR}, Friedrich-Alexander-Universität, Erlangen-Nürnberg

The impact of personalized AI on personal autonomy

The development of LLMs opens the door to personalized AI assistants: AI systems fine-tuned on personal data to replicate the cognitive characteristics of a person, aimed at assisting the person in performing cognitive tasks, with various degrees of independence and proactiveness on the side of the artificial system (see e.g., Earp et al., 2023; Porsdam Mann et al., 2023; Danaher & Nyholm, 2024; Gabriel et al., 2024; Kuilman et al., 2025). In virtue of their replicative and agentic capabilities, personalized AI assistants acting on one's behalf bring about unprecedented implications for personal autonomy. Drawing on Catriona Mackenzie's (2014; 2019) multidimensional and relational framework, this paper analyses the impact of different kinds of personalized AI assistants on three distinct but interrelated dimensions of personal autonomy (self-determination, self-governance, self-authorization). This analysis thus spells out the ways in which personalized AI assistants can either support or undermine the person's opportunities, capacities, and authority in making and enacting one's own choices. In so doing, this analysis also provides initial recommendations for the development and deployment of personalized AI assistants that respect and support personal autonomy. Through an autonomy-based analysis, this paper sheds light on the profound and ambivalent individual and social implications of an emerging AI technology, thus contributing to inform its ethical assessment and public regulation.

Keywords: personal autonomy; personalized AI assistants; LLMs; Mackenzie; AI Ethics

Eleonora Catena is a doctoral student at the Centre for Philosophy and AI Research at Friedrich-Alexander-Universität Erlangen-Nürnberg. Her dissertation examines the impact of AI technology on the concept and exercise of human autonomy. She previously completed an advanced school in AI (ISTC-CNR; AI2Life), after graduating with a Master's in Politics, Philosophy and Public Affairs (University of Milan "La Statale") and a Bachelor's in Philosophy ("Sapienza" University of Rome). Her research interests include various ethical fields (bioethics, animal ethics, AI ethics), as well as moral and political philosophy.

Yalena Chothia
University College London

Technoscientific re-orientations of human connection and the cold intimacies of artificial intelligence

This paper engages with the concept of "cold intimacies", intimate relations refracted through the interiorisation of consumer choice and calculative rationalization. This 'cooling' of emotions results in a cognition of relationality regulated by optimization and transaction, undermining efforts at human connection. By examining how artificial intelligence contributes to the experience of human connection under technoscientific rationality, this paper argues that AI companionship amplifies and legitimize the transference of human relationality into the sphere of consumer leisure. The first section examines the manner in which alienated socialization amplifies the call upon people to congnise relationality on the order of consumption and how the recasting of relations as frictionless consumer experiences makes AI companionship plausible on the affective level. The second section reviews the structural features of LLMs which disrupt consistent

social and emotional integration. The third section focuses on how technoscientific rationality reconfigures the our relation to temporality, interfering with the experiential processes through which intimacies are lived. The fourth section evaluates the principal objections to our thesis. The paper concludes that, despite the alleged benefits for emotional and social wellbeing, with AI companionship the anthropomorphosis of capital becomes a fait-accompli.

Keywords: Affective capitalism; Artificial intelligence; Human connection; Philosophy of technology; Technoscience

Yalena Chothia is a Researcher in Philosophy and STS. She earned a B.Sc. (Hons) in Human Sciences and M.Sc. in History and Philosophy of Science from University College London where she specialised in historical and social epistemology. Yalena's current research involves the interaction between pharmacology and broader culture and how the governance of emerging technologies can meet the interests of people and society. She is also currently preparing her P.h.D. thesis on the importance of silence in mediating our experience of nature. She has presented her M.Sc. thesis « Philosophical and biological conceptions of life » at Universiteit Gent at the third Logic and Life conference in July 2024. She has also participated in the Measuring the Human conference at the University of Cambridge and has two forthcoming articles on meditation as an aesthetic experience and the biopolitical naturalisation of aesthetics. Her academic interests include pragmatism and the philosophy of Charles Sanders Peirce, the history of aesthetics and visual cultures between the eighteenth and twenty-first centuries, and how the ethics and technologies of selfhood inform our relation to the environment.

Julia Egenhoff

Carl von Ossietzky Universität Oldenburg

Digital body, coded femininity: FemTech and the constitution of "corpo-realities" under conditions of digital governance

The lecture "Digital Bodies, Coded Femininity: FemTech and the Constitution of 'Corpo-Realities' under Conditions of Digital Governmentality" builds on work within feminist philosophy of technology and critical AI studies with a focus on digitalization in healthcare. Drawing on soma studies, it combines approaches from post-phenomenology (Merleau-Ponty, Ihde, Hansen) with neo-materialist research (Barad, Haraway) to examine the interdependencies of body, gender, and digitization.

Physical autonomy requires the availability of medical knowledge that enables informed decisions and awareness of one's own body. At the same time, the techno- scientific interpretation of the body is an essential factor for political agency due to its implicit normativity. It not only shapes the physical self-image and the attribution of affordances, but also expands or limits political scope for action through its institutional, structural, and legal materialization and application (Burrow 2021: 126). However, the "gender health gap" shows that the allocation of data about the human body exhibits a gender asymmetry (Dusenberry 2018; Graham 2023). Medical knowledge is primarily based on male-read bodies and reproduces sexist assumptions. According to Ida Tin, founder of the menstruation app Clue, combating gender asymmetry requires technical intervention. In 2016, she called this "FemTech," which includes hardware and software designed to promote women's health (Corbin 2020; Balfour 2023). But is FemTech a liberation movement that subjects medicine to a feminist correction, or does the positivist-technocratic promise of

physical self-enlightenment follow a new form of biopolitical control under conditions of digital governmentality? This is where the lecture comes in. It examines FemTech as technoscientific "agents" of epistemic and physical production in order to answer how FemTech co-constitutes "corpo-realities" under conditions of digital governmentality. Three theses will guide the lecture:

1. Digital FemTech products co-produce gendered bodies through socio-somatic adaptation to their use. From a neo-materialist and post-phenomenological perspective, the use of FemTech products can therefore be understood as a continuous restructuring of bodily dispositions.

2. FemTech products are instruments of digital governmentality that co-constitute bodies along the modes of normalization, functionalization, naturalization/embodiment, and depoliticization.

3. Nevertheless, FemTech products have subversive potential. Through queering, they can become tools for physical reappropriation, political agency, enlightenment, and collective solidarity.

Keywords: Digital Governmentality; AI applications in healthcare; Gender and AI; FemTech; New Materialism

Short bio-bibliographical profile: Julia Egenhoff works as a lecturer in the history of philosophy at Carl von Ossietzky University in Oldenburg. She is currently writing her doctoral thesis on "Digital bodies, coded femininity: FemTech and the constitution of 'corpo-realities' under conditions of digital governmentality" and is conducting parallel research on issues of digital ecology. She is also an editor of the Historical Critical Dictionary of Marxism and is involved in various research groups on topics of digital injustice and colonial continuities.

Emily Elstub

Centre for the Future of Intelligence, University of Cambridge

Safeguarding the mind: against an absolute right to mental integrity

Most rights are relative and thus may be limited to protect public interests or the rights and freedoms of others. A small number, however, are widely regarded as absolute, meaning that infringements can never be justified. Alongside the prohibitions of torture and slavery, the right to freedom of thought – under which the right to mental integrity is widely viewed to be subsumed – has long been considered absolute (e.g., Alegre, 2017; Shaheed, 2021). The implications of this are profound. If mental integrity is absolute, then any limitation, by any means and for any ends, would be prohibited. On this view,the impermissible alteration of an individual's mental states can never be justified. Recent scholarship, however, has called the presumed inviolability of freedom of thought, and by extension mental integrity, into question, proposing instead a relativistic account, albeit constrained by strict conditions on permissible limitations (e.g., Bublitz, 2021, p. 90). Yet both the scope of the protections mental integrity should afford and the normative justification for those protections remain underdeveloped and contested. Further, these debates often rest on an uncertain or unsubstantiated conception of both freedom of thought and mental integrity. This paper defends a broader relativist account, showing that mental integrity admits a wider range of justified limitations than prior approaches have acknowledged. I argue that both intra- and interpersonal limitations may be permissible when necessary to either prevent or restore substantial deficits in mental integrity. Rather than regarding mental integrity as an inviolable barrier, this view presents it as a dynamic safeguard whose protection may sometimes require carefully constrained limitations. By situating mental integrity within this broader normative framework, the paper clarifies its relationship to freedom

of thought, resolves ambiguities within the current scholarship and provides a coherent and principled framework for evaluating interventions aimed at augmenting mental states.

Keywords: Mental, Integrity, Neurotechnology, Absolute, Relative

Emily Elstub is a PhD candidate in philosophy and a student fellow at the Centre for the Future of Intelligenceat the University of Cambridge, and a research associate at the Institute for Ethics in Technology at the Hamburg University of Technology. She received a bachelor's degree in philosophy and psychology from the University of Oxford, and a master's degree in philosophy from Utrecht University. Her research examines the normative and epistemic implications for human rights posed by artificial intelligence and neurotechnologies.

Ryan SangBaek Kim
Ryan Research Institute (RRI)

Affective Sovereignty: Reclaiming the Right to Feel for Oneself

As artificial intelligence evolves from informational tools to relational partners, a new ethical frontier emerges: the erosion of interpretive sovereignty over one's emotional life. Contemporary AI systems—companions, wellness bots, affective computing—increasingly interpret, validate, and anticipate human emotions. When algorithms become the primary interpreters of what we feel and why, they create interpretive displacement: the systematic precedence of algorithmic authority over phenomenological self-access. This paper introduces Affective Sovereignty, defined as the inalienable right to be the ultimate authority on one's emotional experience. We argue this right is foundational to human dignity and agency, yet remains unprotected by existing frameworks. While GDPR addresses data rights and IEEE standards target bias, neither safeguards the right to interpretive primacy—including the right to be uncertain, ambiguous, or mistaken about one's feelings without algorithmic correction.

The philosophical core rests on the non-delegability of affective self-knowledge: emotions are not external facts to be measured, but lived meanings constituted through interpretation. When AI systems bypass this process, they commit a "uniqueness violation"—the expropriation of the hermeneutic space where selfhood is enacted.

To operationalize Affective Sovereignty, we propose a Sovereign-by-Design framework grounded in three principles: (1) Interpretive Restraint—AI must suggest, not diagnose, preserving epistemic humility; (2) Emotional Provenance—Users must transparently distinguish their interpretations from AI-generated ones; and (3) Identity-Responsive Feedback—Users retain the right to reject, correct, and delete AI's affective models. Reframing emotion as a domain of rights, not merely data, provides a human-centered foundation for the ethical governance of affective AI.

Keywords: Affective Sovereignty; Interpretive Rights; AI Ethics; Phenomenology of Emotion; Human-Centered AI

Ryan SangBaek Kim, Ph.D., is a Paris-based transdisciplinary scholar working at the intersection of affective neuroscience, AI ethics, psychoanalysis, and philosophy. Ryan SangBaek Kim is the founder and director of the Ryan Research Institute (RRI), where he leads the Affective Sovereignty Studies program on emotional rights, interpretive ethics, and affective autonomy in human–AI relations. Current research spans

three complementary projects: Affective Sovereignty (ethical and human-rights foundations of emotion AI), Predictive Emotional Selfhood in Artificial Minds (on intrinsic safety and affective alignment, submitted to IASEAI'26, Paris), and Cognitive Circuit Breakers (on epistemic safety and AI governance, submitted to the Wharton AI Governance Conference). The broader aim of this work is to articulate emotional dignity and cognitive autonomy as non-negotiable conditions for human-centered AI.

Petr Jošt
Department of Philosophy and Social Sciences, Philosophical Faculty, University of Hradec Králové

AI Moral Assistant: Machine Ethics in the Context of Moral Enhancement

Machine ethics addresses issues related to the potential development of artificial moral agents (AMAs), that is, autonomous systems capable of complex moral reasoning. Although well-established within AI ethics, the topic has recently been linked to efforts to enhance human morality. Moral enhancement generally refers to attempts to improve moral character or capacities, mainly through biomedical interventions, to prepare our evolutionarily outdated moral psychology for contemporary challenges such as global inequality and climate change. However, effective and reliable biomedical interventions are not yet available, making much of the debate highly speculative. However, recent advances in artificial intelligence have led to what appears, at least initially, to be a more practical proposal to develop AI moral assistants, which I will examine in my talk. In the first part, I will present and define the distinction between substitution and auxiliary approaches. While the substitution approach aims to replace a person's moral reasoning altogether, the auxiliary approach seeks primarily to assist and, ideally, improve the process of moral reasoning. The role of auxiliary systems could, for example, include supporting the selection and processing of relevant information and flagging cognitive biases and inconsistencies throughout the reasoning process. In the second part, I will review empirical evidence on how people use current AI systems, focusing on whether people turn to them for moral advice, what kinds of guidance they request, and whether receiving that guidance affects their moral reasoning or beliefs. I will conclude my talk by examining the near-term feasibility of such systems and whether their creation would be desirable.

Keywords: AI moral assistant, machine ethics, moral enhancement, speculative bioethics, artificial moral agent

My name is Petr Jošt and I am a PhD candidate in the Department of Philosophy and Social Sciences, Philosophical Faculty, University of Hradec Králové. My dissertation examines practicable, currently available avenues of moral enhancement, and this has led me to analyse the capabilities and limits of contemporary AI systems. My broader interests include applied ethics, bioethics and speculative bioethics, with a focus on prediction, regulation and public education on emerging biotechnologies. I am also interested in issues around folk psychology and cognitive ontology, particularly in relation to biomedical moral enhancement, and I have recently submitted an article on this theme titled "Biomedical Moral Enhancement and the Issue of Core Moral Dispositions". Beyond the dissertation, I focus on early modern philosophy, which I teach at the university.

6.
SOCIETÀ E DEMOCRAZIA
SOCIETY AND DEMOCRACY

Contributors:

1. Arianna Atzeni
2. Gianluca Baldassarre
3. Paula Borges Liebana
4. Vincenzo Borrelli
5. Isabella De Vivo
6. Luisa Damiano
7. Marco Fasoli
8. Alice Helliwell
9. Daniel Melbert
10. Fabiana Miraglia
11. Alfred "Win" Mordecai
12. Georgie Newson
13. Sara Pane
14. Vikram Sura
15. Luca Tenneriello
16. Maran Ida Weeber

Arianna Atzeni
Facoltà di Scienze Politiche, Università degli Studi di Roma Tre

Safeguarding Freedom as Non-Domination in the Age of AI: a Neo-Republican perspective on Bias and Political Instrumentalization

The swift progression of artificial intelligence (AI) presents considerable threats of domination, particularly when analysed through the framework of neo-republican theory, which emphasizes freedom as non-domination. This paper explores two critical dimensions of AI's potential to dominate: its inherent biases and its political instrumentalization. First, AI systems, trained on datasets reflecting historical and societal biases, risk perpetuating and amplifying structural inequalities. Such biases can lead to the domination of marginalized groups, as AI-driven decisions in areas like law enforcement, hiring, and healthcare disproportionately affect minorities, undermining their freedom and autonomy. Neo-republican theory highlights the danger of unchecked power, and in this context, AI becomes a tool of domination when its decision-making processes lack transparency, accountability, and inclusivity.

Second, the collaboration between governments and technology corporations raises concerns about the political use of AI to dominate populations. Agreements between states and tech giants often prioritize efficiency and control over individual freedoms, creating systems of surveillance and manipulation that threaten democratic principles. Neo-republican theory critiques such concentrations of power, arguing that they enable arbitrary interference in citizens' lives. The fusion of state authority and corporate AI capabilities risks entrenching a new form of domination, where citizens are subjected to opaque algorithms and data-driven governance.

To mitigate these risks, this paper advocates for robust regulatory frameworks grounded in neo-republican principles, ensuring that AI development and deployment prioritize non-domination, transparency, and equitable outcomes. By addressing both the biases embedded in AI systems and the political alliances that exploit them, society can harness AI's potential while safeguarding freedom and justice for all.

Keywords**:** neo-republican theory, freedom as non-domination, power imbalance, AI bias, tech-oligarchy

Arianna Atzeni: PhD candidate in Political and Moral Philosophy at Università degli Studi "Roma Tre". I graduated in March 2024 with a thesis entitled: "Preserving our Planet. Green Republicanism and Environmental Virtue Ethics for a Climate Change Response". I am particularly interested in exploring different contemporary issues from both an ethical and a socio-political point of view. In particular, I am interested in green republicanism; politics, economics, ethics of sustainability; environmental justice, concrete utopias, environmental virtue ethics, power dynamics and dominion, human-animal studies; AI ethics and bioethics.

Gianluca Baldassarre

Istituto di Scienze e Tecnologie della Cognizione, CNR

The most important effects of artificial intelligence will be...

I define the importance of an effect of a given event as the difference in value produced by that effect within a specific time horizon, relative to a set of values and goals held by an evaluating agent.

Based on this concept, I pose the following question: What will be the most important effect that artificial intelligence (AI) will have within the next 15 to 30 years? To address this question, I consider humanity as a whole as the evaluating agent. I then identify the most important effects according to three criteria implicit in the definition:

1. Which elements of society will be most strongly impacted by AI?
2. Which of these elements hold the highest value for humanity?

These criteria lead me to identify two main candidate effects:

1. AI generates enormous amounts of economic wealth due to its self-generating and self-sustaining capacity;
2. This, combined with the positive feedback mechanisms intrinsic to economic systems, leads to massive concentrations of power in the hands of a few companies and states.

Such dynamics may result in increased inequality and, ultimately, a net negative value for humanity. I then address a second question: What could be done to counteract this tendency? The primary solutions appear to be:

1. Strengthening the connection between the evaluating agent carrying the interest of overall well-being–humanity as a whole– and political power;
2. Ensuring that political power acts to distribute economic power rather than allow its concentration.

I close on a pessimistic note, as I believe humans are largely unprepared to understand, desire, and implement these types of actions.I therefore urge immediate efforts to increase such preparedness.

Keywords: AI, economic wealth, concentration of power, political power, division of power.

Gianluca Baldassarre received a bachelor and master in economics in 1998, and a Specialization Course in "Cognitive Psychology and Neural Networks" in 1999 from the Sapienza University of Rome, Italy; and a PhD in Computer Science in 2001 from the University of Essex, Colchester, U.K. He then joined the Italian Institute of Cognitive Sciences and Technologies, National Research Council, in Rome. Here he was: Post-doc since 2003; Researcher since 2006; Coordinator of the Research Group Laboratory of Embodied Natural and Artificial Intelligence since 2016; Director of Research since 2021. He was also Coordinator of the EU Projects IM-CleVeR – Intrinsically Motivated Cumulative Learning Versatile Robots, 2009-2013; GOAL-Robots – Goal-based Open-ended Autonomous Learning Robots, 2016-2021; General-purpose robot for object retrieval in warehouses, 2021-2023. He was the founder and President of the Advanced School in AI in 2018-2025. His research interests span open-ended learning of sensorimotor skills, extrinsic and intrinsic motivations, and goal-directed 95ocused95 –in animals, humans, and robots. He has more than 100 peer review publications on these topics.

Paula Borges Liebana

Universitat Pompeu Fabra – BIAP Maria de Maeztu Fellow

Beyond the Fragmentation Debate: Understanding Social Media's Public Sphere through Epistemic Enclosure

This paper examines the transformation of the public sphere in the wake of AI-driven social media, situating its effects within the framework of deliberative democratic theory. The rise of algorithmic recommender systems marks an "artificial turn" in social media, fundamentally altering the epistemic conditions under which public discourse unfolds. While traditional debates on digital fragmentation focus on polarization, this paper argues that they fail to capture the structural logic of AI-driven platforms, which shape public discourse through mechanisms of epistemic enclosure. Unlike theories of echo chambers and filter bubbles, epistemic enclosure accounts for the systemic constraints imposed by platform architectures, economic imperatives, and the logic of engagement-driven content curation. In this way, epistemic enclosure captures a subtler but more profound transformation: AI systems do not merely distort information but condition the terms under which discourse and political claims can emerge.

By prioritizing attention-maximization over democratic deliberation, AI-driven social media reconfigures the boundaries of political discourse, influencing not just what individuals see, but what can be collectively known and contested. This process consolidates asymmetrical epistemic power, where corporate and state actors shape the parameters of public reason without direct coercion. In contrast to the ideal of an open and contestatory public sphere, AI-driven mediation introduces a paradox: while increasing access to information, it simultaneously narrows the discursive field through predictive filtering, automated moderation, and engagement-driven ranking. This paper calls for a critical reassessment of the public sphere in an era where artificial intelligence functions as an invisible gatekeeper of political visibility and epistemic legitimacy, emphasizing the need for structural interventions to safeguard pluralistic contestation and democratic agency in the digital public sphere.

Keywords: Social Media, Fragmentation Debate, Epistemic Enclosure, AI-Driven Public Sphere

Paula holds a Bachelor's in Politics, Philosophy, and Economics (Goldsmiths, University of London), an MSc in Media and Communications (LSE), and an MA in Philosophy for Contemporary Challenges (UOC).

Currently, she is pursuing a PhD in Political Philosophy at Pompeu Fabra University, exploring the transformation of the public sphere under the revolution of data-driven AI within social media platforms. Her research is funded by the Barcelona Institute of Analytic Philosophy (BIAP), a Maria de Maetzu Unit of Excellence financed by the Spanish Ministry of Science and Innovation.

Vincenzo Borrelli
Certiquality

Governance and Management Systems of Artificial Intelligence. Human-centered models, voluntary regulations, and soft law for AI

The application of AI in businesses introduces innovation and transformation in operational management and can be used to improve and streamline decision-making, governance, and managerial processes in relation to the strategic and value objectives of the business (both for-profit and non-profit). AI applications are subject to binding legislation, which has been developed in various ways in recent years and is still evolving in some respects, aimed primarily at determining the scope of use and "pre-ordained" classification of the associated risks.

In the practical application of new technologies, and therefore AI, organizations are increasingly orienting their approaches according to corporate policies and codes based on values and contexts that can change rapidly. In this scenario, alongside the legislative framework that has been consolidated at international and national level, a framework of voluntary regulation and soft law is emerging at international level that responds to the demand from modern organizations to determine objectives and policies relating to the use of AI not with a prescriptive or binding approach, but through risk/opportunity management models based on the ability to address and make decisions on issues relating to reliability, fairness, security, transparency, and data quality throughout the entire AI lifecycle, including the management of suppliers, partners, and third parties that develop AI systems for the organization itself. Management systems for artificial intelligence, based on the principles of continuous improvement characteristic of every risk-thinking-based approach, can offer a systemic response to those elements and characteristics of AI use that cannot be based solely on compliance with legislative and/or regulatory obligations and which are obviously not adequately regulated by recent legislation. Alongside the prescriptive component and the risk of sanctions, it is necessary to seek methods and approaches that ensure that AI applications, in addition to being in compliance with regulations, are able to effectively safeguard the expectations of users and stakeholders in terms of risk/opportunity. Similarly, therefore, to other sectors of the economy, the International Organization for Standardization (ISO) has introduced a set of voluntary standards on AI for this purpose, starting in 2023, including ISO/IEC 42001:2023 and ISO/IEC 23894, which provide an integrated model for the responsible use of AI and the assessment of impacts for the responsible pursuit of its objectives in a human-centered approach.

Keywords: Governance e risk management, Voluntary AI standardization, Human-centered models, ISO standards for AI, AI life cycle

Luisa Damiano
IULM University

"Robots with Us, Not Like Us": An Epistemological, Ethical, and Design Paradigm for Social Robotics

This paper presents the paradigm "robots with us, not like us", developed through transdisciplinary collaboration between philosophy of science, social robotics, and art, and implemented in ScentDia (2025), the first robot featuring olfactory social presence.

The paradigm's epistemological core lies in a relational view of sociality that integrates insights from cybernetic, autopoietic, and enactive epistemologies. In this view, sociality is not an intrinsic property of individual agents but an emergent process of inter-individual coordination. On this basis, social robots are conceived as new social actors: artificial agents capable of participating in dynamics of social signaling with humans through synthetic mechanisms of social coordination that do not reproduce the biological foundations of human sociality. From this epistemological ground arises the ethical stance defining the paradigm. The traditional ideal of imitation in social robotics—introduced to facilitate interaction but blurring human and robotic identities—is replaced by the ideal of compatible difference, which values difference and complementarity. Ethical responsibility thus shifts from reproducing human likeness to designing forms of artificial otherness that are perceptible, intelligible, and socially engaging. This ethical positioning grounds the design notion of the perceptual identity marker: a sensory and social cue that makes robotic alterity perceptible and facilitates interaction. The first implementation of this principle is ScentDia, an artistic social robot developed in collaboration with artist and roboticist Mari Velonaki and perfume creator Manos Gerakinis. ScentDia features a unique, non-biological olfactory signature functioning as a perceptual identity marker—a perceptual and social signal that makes its difference immediately legible and activates coordination with humans. With ScentDia, olfaction—an evolutionarily ancient medium of social communication—becomes a new path for relational and ethically sustainable social robotics.

The talk will present the paradigm and the design process of ScentDia, discussing their implications for the ongoing diffusion of social robots.

Keywords: olfactory social robotics; perceptual identity marker; olfactory social presence; philosophy of social robotics; ScentDia

Luisa Damiano is Full Professor of Logic and Philosophy of Science at IULM University (Milan). Her research explores the epistemological and ethical dimensions of the sciences of the artificial, with a focus on synthetic biology and on cognitive and social robotics. On the topic of social robots, she has published numerous scientific articles and the book Living with Robots (with P. Dumouchel, Harvard University Press, 2017), originally written in French and also published in Korean and Italian.

Isabella de Vivo
Sapienza Università di Roma

La sovranità della decisione umana tra presunzione di affidabilità e standardizzazione tecnica dei sistemi di AI: limiti e prospettive del Trustworthy Ecosystem delineato dall'AI Act

Obiettivo di questo contributo è analizzare se, e in quale misura, il "Trustworthy Ecosystem" delineato dall'AI Act riesca a garantire, oltre che a "comunicare", il rispetto del principio cardine alla base della visione umano-centrica di affidabilità dell'AI di cui alle Linee Guida 2019 della Commissione europea: la sovranità della decisione umana nella relazione machine-human, intesa (anche) come autonomia decisionale dei soggetti impattati dall'utilizzo di determinate tecnologie (Human Recipients – HRs) nella definizione dell'an e del quomodo dello sviluppo di determinati sistemi di AI, in specifici contesti d'uso. In particolare, l'analisi verterà sulle criticità emergenti dalla declinazione "espertocratica" di AI affidabile delineata dal Legislatore europeo e che vede l'accentramento dell'autorità decisionale negli organismi di standardizzazione tecnica e indirettamente negli AI providers. Il giudizio di conformità tecnica crea, infatti, una presunzione de iure di accettabilità politica dei rischi riguardanti la categoria immateriale dei diritti fondamentali da cui deriva, in via automatica, l'"etichetta" di affidabilità di un sistema di AI. Si rifletterà, dunque, sulle possibilità offerte da un approccio ecologico, o community-based: un modello decisionale complementare che mira a collegare la progettazione e l'implementazione dei sistemi con norme, valori e standard specifici della "comunità ospitante", lungo l'intero ciclo di vita dell'AI. Un quadro orizzontale che, facendo perno sulla nozione di fiducia come processo negoziato e iterativo, sia in grado di recuperare il coinvolgimento democratico degli HRs nella definizione, in concreto, dell'ammissibilità delle soglie di rischio poste ai diritti fondamentali dai sistemi con cui, volontariamente o meno, saranno destinati a interagire. Si argomenterà, in conclusione, come la denaturalizzazione della visione dell'AI come prodotto che la sovrapposizione tra standardizzazione tecnica e (trust)worthiness presuppone, sia precondizione per limitare l'autorità epistemica delle big tech, quali arbitri de facto del bilanciamento dei diritti fondamentali coinvolti nello sviluppo e implementazione dei loro sistemi e modelli di AI.

*

Human Decision Sovereignty, Trustworthiness and Standardisation under the EU AI Act

To what extent does the EU AI Act's "Trustworthy Ecosystem" truly succeed in safeguarding human decision sovereignty—the very core of the European Commission's human-centric vision of trustworthy AI?

This analysis highlights the ethical and political tensions embedded in the "expertocratic" model of trustworthy AI set up by the EU AI Act. Trust, in fact, is a bidirectional process between trustor and trustee and, as such, presupposes the decisional autonomy of Human Recipients (HRs) as primary trustors. Yet, the AI Act largely concentrates decision-making authority over the assessment of trustworthiness in technical standardisation bodies—and, indirectly, in AI providers—whose decision-making procedures structurally exclude democratic participation.

Technical conformity assessments thus generate a de iure presumption of acceptability of risks affecting the immaterial sphere of fundamental rights, from which the "trustworthy" label of AI systems automatically derives. This shift risks neutralising the social and relational dimension of trust by displacing value-laden judgments from democratic deliberation to technocratic procedures.

Against this backdrop, the contribution reflects on a community-based approach to trustworthy AI as a complementary decision-making model. By reconnecting AI system design and deployment with the norms, values, and standards of the "host community" throughout the entire AI lifecycle, such a model seeks to restore the democratic involvement of Human Recipients in determining acceptable thresholds of risk to fundamental rights.

This perspective highlights that denaturalising the conception of AI as a product resulting from the overlap between technical standardisation and (trust)worthiness constitutes the first step in limiting the epistemic authority of AI providers and in moving towards a genuinely human-centred conception of AI.

Keywords: Human Decision Sovereignty; HLEG's Human-Centric Trustworthiness; AI Act Trustworthy Framework; Community-based Approach.

Isabella de Vivo, avvocato, laureata in Giurisprudenza cum laude presso l'Università Luiss G. Carli di Roma, ha successivamente conseguito la laurea specialistica in Lettere e Filosofia all'Università Sapienza di Roma, con votazione 110 e lode. Presso la medesima Facoltà ha conseguito il Dottorato di ricerca in Storia e Culture dell'Europa (Dipartimento SARAS). È stata visiting researcher presso l'Institute for Information Law (IviR) dell'Università di Amsterdam (UvA) con un progetto dal titolo "Algorithm as a Social System: A Model for Investigation". È cultrice della materia per i corsi di Diritto Pubblico di Internet e Istituzioni di Diritto Pubblico presso il Dipartimento DEIM dell'Università degli Studi della Tuscia, dove ha svolto attività di docenza integrativa. I suoi interessi di ricerca, intersecando la prospettiva giuridica e filosofica, vertono sull'impatto dell'Intelligenza Artificiale sui diritti fondamentali, con particolare focus sul diritto alla protezione dei dati e all'identità personale, libertà di espressione, non discriminazione, governance algoritmica, governance delle piattaforme e delle infrastrutture digitali, muovendo dall'emergente regolazione europea della materia. È attualmente titolare della borsa di ricerca dal titolo "Identificazione biometrica, privacy ed etica del servizio sociale" nell'ambito del progetto "ALOHA, a framework for monitoring the physical and psychological health of workers through object detection and federated machine learning", presso l'Università degli Studi del Molise.

Marco Fasoli

Sapienza Università di Roma

The non neutrality of generative A.I.

"A.I. is neutral, it is neither good or bad, it has no ends of its own and exists only to accomplish human ends". This common-sense belief is often mentioned in the public debate, in order to attribute the entire responsibility of some human actions to the people. In the philosophy of technology, similar claims are usually defended by the so called "technological instrumentalism" (Pitt 2014). According to technological instrumentalism, technology is simply a mean and we are totally free to use any way we want. Conversely, according to "technological determinism" technology somehow determines how we use it. In this paper, I will analyze the starting claim and I will point out that it actually conflates two diPerent theses. The first thesis is that technology cannot embody values, while the second one is that how an artifact is built does not influence how we employ it. Insofar as we usually don't' confer values to physical objects but to actions, choices, thoughts, etc., I will argue that the first thesis seems sound. Conferring values to objects would require the development of a counterintuitive notion of value, but we actually don't have some reasons to do this. On the other hand, I will point out that the second thesis is based on a I conception of the human

rationality. In order to understand how technology may foster some behaviors, we need to assume a more sophisticated model of the human mind. In the last thirty years, cognitive sciences deeply revised our intuitions about the functioning of our mind, identifying a large spectrum of biases (see Kahneman 2011). Furthermore, artifacts have aPordances that oPer us specific repertoires of actions (Casati 2017, Fasoli 2018, Latour 1994). This change of paradigm looks similar to the transition from the classical economy to the behavioral economy (Thaler & Sunstein 1999). I will argue it would also have similar consequences, helping us to achieve a deeper and more realistic understanding of GenA.I.

Keywords: Generative AI, neutrality, bias.

Marco Fasoli è stato ricercatore (RTDA) presso il Dipartimento di Filosofia della Sapienza Università di Roma e attualmente è docente a contratto. Si occupa di filosofia della tecnologia delle scienze cognitive, in particolare dei cosiddetti artefatti cognitivi e dell'impatto che le nuove tecnologie digitali e l'IA hanno sul nostro benessere personale. Attualmente insegna 2 corsi (Mente e azione e Filosofia della I.A.) nel corso di laurea in "Filosofia e intelligenza artificiale" (Sapienza). Ha insegnato nel dottorato in "Sviluppo sostenibile e cambiamento climatico" (IUSS Pavia). Ha pubblicato articoli scientifici in diverse riviste nazionali e internazionali (tra cui Philosophy and Technology, Aphex, Minds and Machines, Sistemi Intelligenti, The Review of Philosophy and Psychology). Nel 2019 ha pubblicato per Il Mulino la monografia "Il benessere digitale" e ha ricevuto il Premio Vittorio Girotto dall'Associazione Italiana di Scienze Cognitive (AISC) per l'articolo "Contro lo strumentalismo tecnologico. Per una teoria analitica della prescrittività degli artefatti".

Alice Helliwell

Northeastern University London

The Role of Convincingness in AI Propaganda

AI has been described as a "disinformation machine" (Stanford HAI, 2024), able to rapidly generate content that can be used to mislead, persuade, and manipulate. Recent studies indicate that "that propagandists could use AI to create convincing content with limited effort." (Goldstein et al. 2024) and that "generative-AI tools have already begun to alter the size and scope of state-backed propaganda campaigns." (Wack et al. 2025). This does not just extend to the use of AI generated text and images, but increasingly video and audio.

In this paper I will examine some recent examples of AI use in propaganda. I will then go on to present a dimension of generative AI outputs that I call 'convincingness'. Convincingness goes beyond typical conceptions of media realism. It concerns not whether a photographic image (for example) convinces us that what it depicts actually happened, but rather that it is a photograph at all. The 'convincing' has been less of a concern for traditional media. However, with the advent of rapid synthetic media generation with AI, convincingness is increasingly relevant. We might, for example, be convinced that a painting is truly a painting (when it is in fact AI generated).

I argue that this feature of convincingness is of particular relevance when discussing AI as a propaganda tool. Coupled with speed and flexibility, the ability of AI to produce outputs that convince the casual viewer that they are of a certain medium adds to their potency. I also argue that the convincingness of AI generated propaganda is distinct from its persuasiveness, though both are relevant dimensions of AI-generated propaganda, and may be correlated.

Keywords: Artificial Intelligence, Generative AI, Propaganda, Aesthetics, Ethics

Alice Helliwell is an Assistant Professor in Philosophy at Northeastern University London. Alice's research is 101ocused on AI art and computational creativity. She holds a PhD in the History and Philosophy of Art from the University of Kent. Alice's research focuses on machine creativity, the interaction between art and AI, and the aesthetics of AI images. She has also published on AI ethics, Wittgenstein and AI, and AI value alignment.

Daniel Melbert
Purdue University

Being, Alienation, and Production in the Age of Generative Artificial Intelligence

With a continuous influx of investment and large-scale adoption by institutions, Generative AI tools—most notably Large Language Models—have grown to be present across nearly every facet of contemporary life. This presence has been met with growing social concerns surrounding increasing social isolation, potential for mass unemployment, misinformation, environmental impact, and encoded bias. The intertwined history of capital, mechanization, and its influence in the deployment of industrial machinery across history, suggests the necessity for an updated Marxist analysis of the emergent technologies redefining contemporary life. Marx's alienation has gained a resurgence in consideration within the present as it emphasizes the impact of labor organization upon workers and society at large.

While scholars have succeeded thus far in reasserting the importance of alienation across Marx's work and within the present, such accounts remain limited by the focus of Marx—that being the capitalist organization of production. These accounts stand to be strengthened by extending consideration to the nature of production and human activity offered by a decidedly more technological focus. Within this paper I analyze the nature of technological production engaged in by humans from Marxist, Kappian, and Ruyerian lenses. I bridge these perspectives to construct an account of alienation that supplements the underdeveloped ideas across Marx's account. In doing so, I contextualize the unique mechanism of estrangement generative AI technologies impose upon the subject—in which the gradual cognitive offloading of neural processes onto machines impedes engagement in critical thought and creative action. This extends the nature of alienation beyond the labor act, into a mechanism of logical encoding and epistemic imposition. This mechanism offers a potential frame of future inquiry in AI, production, capitalism, technicity, and creativity.

Keywords: Production, Human, Alienation, Activity, Capitalism

Daniel Melbert is a second-year Master's of History student at Purdue University in West Lafayette, Indiana. He graduated this past May with Bachelor's of Arts degrees in History and Philosophy, the latter holding a focus on the philosophy of AI. He currently works as a research assistant in the Purdue Normativity and Cognitions (PuNCs) lab. Guided by a longstanding interest in technology, history, and philosophy, alongside passions for human rights, decolonization, and data ethics, his historical research interests lie in the intersections of technology, class, colonialism, mass movements and the experiences of historically marginalized and oppressed peoples. His philosophical research interests lie in philosophy of technology, creativity, epistemology, critical theory, philosophy of action, and philosophy of solidarity, alongside particular interests in Marx, Foucault, Ruyer, and Deleuze.

Fabiana Miraglia
Università del Salento

Verso un'Intelligenza Spirituale Artificiale (ISA): per una *governance* semantica e interculturale dei modelli linguistici

Il contributo propone una riconcettualizzazione dell'intelligenza artificiale in prospettiva antropologico–culturale, attraverso l'introduzione del paradigma dell'Intelligenza Spirituale Artificiale (ISA), intesa non come tecnologia religiosa o mistica, bensì come orizzonte epistemologico e assiologico capace di integrare la dimensione simbolica, valoriale e relazionale dell'esperienza umana nei processi di generazione automatica del linguaggio. L'ISA si colloca tra l'intelligenza cognitiva e quella etica, configurandosi come livello superiore di consapevolezza algoritmica, fondato sull'idea che ogni forma di elaborazione linguistica automatizzata incorpori presupposti culturali, orientamenti morali e specifiche morfologie del senso. In tale prospettiva, la macchina linguistica è considerata un attore epistemico e simbolico, capace di co-costruire significati e di influenzare le rappresentazioni collettive della realtà.

All'interno di questa cornice teorica, il contributo propone il metodo estro–decoloniale (MED) come strumento sperimentale per la decostruzione critica e l'auditing culturale dei modelli linguistici automatici. Il MED, di natura qualitativa e comparativa, mira a interrogare le infrastrutture semantiche dei sistemi di *Natural Language Processing* (NLP) attraverso una lente etico–culturale. Esso si articola in tre fasi interdipendenti: diagnosi semantica, per individuare i frame interpretativi impliciti; estroflessione concettuale, per ricondurre i concetti generati alle loro genealogie culturali e simboliche; *backpropagation* normogenetica, per tradurre l'errore culturale in apprendimento riflessivo e ricalibratura semantica.

A sostegno operativo del MED, il contributo propone una griglia di valutazione etico–culturale volta a fornire parametri qualitativi di giustizia epistemica, rappresentatività simbolica e inclusione interculturale. In conclusione, l'ISA si configura come paradigma emergente di *governance semantica* e *ecologia cognitiva*, orientato a educare l'ecosistema linguistico–tecnologico all'ascolto, alla coabitazione e alla negoziazione del senso, superando l'etnocentrismo invisibile che ancora pervade i modelli globali.

Keywords: pluralismo culturale; etnocentrismo semantico; giustizia epistemica; bias

*

This paper reconceptualizes artificial intelligence from an anthropological–cultural perspective by introducing the paradigm of Artificial Spiritual Intelligence (ASI), understood not as a religious or mystical technology but as an epistemological and axiological horizon capable of integrating the symbolic, value-based, and relational dimensions of human experience into automated language generation. ASI stands between cognitive and ethical intelligence, representing a higher level of algorithmic awareness grounded in the idea that every form of linguistic processing implicitly embodies cultural assumptions, moral orientations, and specific morphologies of meaning. In this perspective, the linguistic machine is conceived not as a purely technical device but as an epistemic and symbolic actor able to co-construct meanings and influence collective representations of reality.

Within this framework, the paper proposes the Extro–Decolonial Method (EDM) as an experimental tool for the critical deconstruction and cultural auditing of linguistic models. Qualitative and comparative in nature, the EDM interrogates the semantic infrastructures of *Natural Language Processing* (NLP) systems through an ethical–cultural lens. It consists of three interdependent analytical phases: semantic diagnosis, aimed at identifying the implicit interpretive frames that guide automated responses; conceptual

extroflexion, which reconnects generated concepts to their cultural and symbolic genealogies; and normogenetic backpropagation, which transforms cultural error into reflective learning and semantic recalibration.

As an operational complement to the method, the paper proposes an ethical–cultural evaluation grid designed to provide qualitative parameters of epistemic justice, symbolic representativeness, and intercultural inclusion. In conclusion, Artificial Spiritual Intelligence (ASI) emerges as a paradigm of *semantic governance* and *cognitive ecology*, oriented toward educating the linguistic–technological ecosystem to listening, coexistence, and negotiation of meaning, and toward overcoming the invisible ethnocentrism that still pervades global language models.

Keywords: cultural pluralism; epistemic justice; semantic ethnocentrism; bias

Fabiana Miraglia ha conseguito la laurea magistrale in Giurisprudenza e il dottorato di ricerca (PhD) in Diritti, Economie e Culture del Mediterraneo (SSD IUS/11 – Diritto ecclesiastico) presso l'Università degli Studi di Bari "Aldo Moro", Dipartimento Jonico in Sistemi Giuridici ed Economici del Mediterraneo: società, ambiente, culture. È attualmente cultrice della materia in Diritto ecclesiastico e Diritto interculturale presso il medesimo Ateneo. È stata assegnista di ricerca presso l'Università del Salento, Dipartimento di Studi Umanistici, fino al 31 dicembre 2025, nell'ambito del PRIN 2022 "Migrants, Institutions, Translations – Easy-read Law". La sua attività di ricerca si è concentrata sulla semplificazione e traduzione dei documenti e degli atti giuridici attraverso la metodologia giuridica interculturale e una comunicazione dialogica e cooperativa tra le diversità culturali e religiose, con particolare attenzione all'analisi dei documenti richiesti al migrante dalla pubblica amministrazione e ai rapporti giuridici funzionali alla tutela dei bisogni essenziali (identità personale, nascita, residenza, stato di famiglia, accesso all'abitazione e al lavoro).

Alfred "Win" Mordecai
University of Wisconsin-Milwaukee

The Limits of AI Metacognition and the Consequences for Democracy

This paper explains why AI struggles to exercise metacognition, or the capacity to think about its own thinking, and why this poses a systemic threat to democracy. Metacognition enables agents to recognize the limits of their knowledge and identify cognitive errors; it is the precondition of epistemic agency. I argue that democratic participation requires epistemic agency, since responsible deliberation and voting depend on autonomous belief-formation. Yet even with recent improvements in AI technology, AI still performs poorly at metacognitive reasoning tasks. I examine the computer science that explains why AI fails to exhibit metacognitive reasoning: the types of algorithms that would let a system assess its own limits require unrealistic computational resources. This constraint explains why large language models (LLMs) often fabricate information (hallucinate). When we start trusting AI output as epistemically significant, we begin to rely on information that has undergone no process of self-scrutiny. Democracies cannot function when their citizens form political beliefs on the kind of groundless information that AI produces. I end by suggesting two ways to limit AI's damage to democracy: (i) using redundant systems architecture to minimize AI hallucinations and (ii) keeping our use of AI confined to fixed, well-defined problem spaces.

Keywords: Artificial Intelligence, Metacognition, Epistemic Agency, Political Epistemology, Computational Complexity

Alfred "Win" Mordecai is a graduate student in philosophy at the University of Wisconsin–Milwaukee, and he holds a bachelor's degree in computer science from Duke University. His research focuses on the democratic implications of the AI revolution, and he is particularly concerned with the ways in which AI-generated information will affect our shared information landscape. He is currently investigating how the computational complexity of metacognitive functions in AI systems prevents artificial agents from contributing meaningfully to democratic and epistemic institutions.

Georgie Newson
University of Edinburgh

Id Engines: Generative Artificial Intelligence and/as the 'Collective Unconscious'

Generative AI technologies (GAIs) promise to revolutionise the information landscape. Increasingly, these systems are described by their creators and users via a framework of *collective cognition.* Blaise Agüera y Arcas and James Manyika, senior researchers at Google, write that an LLM constitutes a 'super-intelligence with far greater breadth and average depth than any single person'. Dario Amodei, Chief Executive of Anthropic, describes his company's model as 'a country of geniuses in a data centre.' Eleni Vasilaki, a computational neuroscientist at the University of Sheffield, recently proposed that LLMs should be thought of as instantiations of 'collective human knowledge' (Vasilaki 2025). Such images may seem like mere metaphors, or else sensationalist vehicles for 'AI hype' (Kotliar 2025). But are there reasons to take them seriously? What ethical challenges might arise if they prove legitimate? And if they do not, then what ideological function might they serve?

In this paper, I will argue that the image of GAIs as collective intelligences captures an important feature of these systems: that they are better thought of as novel 'cultural technologies' (Gopnik 2022) than unified, autonomous agents. Moreover, GAIs can 'mirror' (Vallor 2024) socially encoded cognitive structures and symbolic repertoires that might not be identifiable at the individual level – including, most infamously, 'hidden' oppressive biases (Pan et al, 2025). For these reasons, LLMs might plausibly be regarded as akin to a collective Freudian-Lacanian *un*conscious (Zupančič 2025). However, both the 'collective intelligence' and 'collective unconscious' images serve two hazardous functions: *naturalization* and *democratization.* Naturalization, because they depict GAIs as organically emergent syntheses of the data on which they are trained. Democratization, because they imply that consulting a GAI may be similar to approximating a consensus or tapping in to the 'wisdom of crowds'. The language of collectivity in GAI contexts can thus easily serve worrisome political ends, legitimizing the automatization of collective human decision-making processes on egalitarian and pragmatic grounds. As a result, I will argue, we should tentatively resist such framings, despite their conceptual plausibility.

Keywords: Generative AI, LLMs, collective unconscious, collective intelligence, philosophy of mind

Georgie Newson is a PhD candidate at the University of Edinburgh. She is currently working on a political history of the concept of the 'hivemind'. Her recent publications include 'A Humanistic Despair: Interwar Philosophy Between Fascism and Extinction', *International Journal of Philosophical Studies* (forthcoming) and 'To Desire What is Nothing: Simone Weil, Asceticism, and Psychoanalysis', Critical Quarterly 67.1 (April 2025).

Sara Pane
Sapienza Università di Roma
Università degli studi di Torino

Generative AI, Agentic State and Quasi-Otherness: Configurations of Post-Digitality in Contemporary Europe

This contribution develops a theoretical reflection on the concept of post-digitality, proposing an alternative interpretation and defining it as an emerging socio-technical paradigm for understanding the transformations generated by the dynamic interaction between society, politics, and technology in the era of generative artificial intelligence and the future scenarios opened by agentic AI. Starting from the theoretical framework and empirical results of the doctoral research—focused on the integration of AI into the public communication of the European Union through experiments with institutional chatbots—the study, based on a qualitative design combining in-depth interviews (n=74), participant observation, and policy analysis, explores the emerging relationship between human and non-human agency, questioning the role of artificial intelligence as a communicative actor and producer of meaning. The analysis reveals a hybrid condition in which communicative action is distributed among human subjects, generative tools, and decision-automation systems, within a context marked by the growing integration of generative AI. The most recent literature adds another dimension to this framework: agentic AI, which opens up the future perspective of the agentic state—conceived as a new interpretative horizon of public governance characterized by unprecedented relationships between human beings and intelligent systems. This perspective makes it possible to understand how artificial intelligences, conceived in a plural and relational sense, are progressively redefining the logics of production, legitimation, and circulation of public discourse—as well as democratic legitimacy—giving rise to a form of "quasi-otherness" in which technology tends to be configured as a communicating subject rather than a mere technical object. The work argues that post-digitality does not end in the technological dimension but constitutes an epistemic regime in which authenticity, authority, and communicative legitimacy are redefined. This transformation deeply affects the relationship between citizens and institutions, introducing new dynamics of trust, transparency, and participation in the European public sphere.

Keywords: public communication; governance; artificial intelligence; public sector; European Union.

Sara Pane is a PhD candidate in the Joint International Doctorate in Social Representations, Culture and Communicationat the Department of Communication and Social Research, Sapienza University of Rome. She is also a member of the teaching staff and engaged in research activities at the University of Turin, within the Jean Monnet Chair Com4T.EU. Her doctoral research offers an innovative interpretation of the concept of post-digitality, analyzing it in relation to the role of artificial intelligence in the public sector. In particular, her work explores how the adoption of AI contributes to redefining the relationship between citizens and institutions, generating new forms of participation and transformation in the public sphere, with a specific focus on the European context. She is part of the Knowledge Community of the World Health Organization (WHO) dedicated to the development of artificial intelligence in support of society, contributing to interdisciplinary reflection on its ethical, social, and political impacts. She has produced essays and scientific contributions on these topics for academic and research institutions and is co-author of the volume "La rivoluzione socio-tecnica dell'IA: un'analisi sociologica" (Edizioni Nuova Cultura, in press).

Vikram Sura

United Nations Secretariat, New York

Geopolitics and AI Governance: Reconciling Responsible Development, Equity and Power

This paper proposes an examination of the governance tension at the heart of responsible AI development, focusing first within international organizations on their aversion to recommending 'regulation'. The term regulation is viewed with dread, implying a top-down control deemed impractical due to the nature of AI technologies, the extraterritorial market operations of technology companies and the fear of international bodies being seen as empowering governments that seek restrictive regulation. This leaves international bodies caught in a conflict between recommending 'regulation' of powerful technology platforms that offer opportunities but amplify systemic risks and preserving freedom of expression.

This governance dilemma is exacerbated by geopolitical and geoeconomic competition. The United Nations has adopted two AI resolutions, both reflecting broad consensus, but sponsored by the United States and China, respectively. The United States emphasizes market-driven innovation and voluntary guidelines, while China a state-led approach focusing on capacity building. This dual sponsorship of competing visions risks effective governance that keeps people's welfare at its core. The European Union A.I. Act with guardrails against degrees of risks from AI systems has a potential to emerge as a third pole. Nonetheless, ethical imperatives guiding responsible global AI governance are not completely absent. A multilateral policy agenda exists in the Global Digital Compact, in the formal establishment of an annual global dialogue on AI to discuss and review timely inputs from an international independent scientific panel on AI, structured along IPCC lines.

Yet the critical challenge is operationalizing these commitments amid geopolitical and geoeconomic competition, where states and tech firms are incentivized toward regulatory arbitrage. When market dominance and strategic advantage eclipse equity, even strong frameworks risk being performative. Without addressing inequities that let some nations wield massive material power to dominate the AI stack while others remain consumers of foreign-trained models and agenda setting, no ethical governance can be truly legitimate or enforceable. Is the community of nations and civil society up for this task, or have they unwittingly yielded to market competition and geopolitical rivalry, risking fragmented AI governance?

Keywords: AI governance, regulation, geopolitics, equity.

Vikram Sura is a United Nations officer since 2001. His work on information integrity involves examining geopolitical influences on the development of governance frameworks that prioritize human welfare as a core tenet of AI development. The views are the author's own and do not represent those of his employer.

Luca Tenneriello

Dipartimento di Ingegneria Informatica, Automatica e Gestionale "A. Ruberti", Sapienza Università di Roma

The Trolley Problem in the Age of AI: Ethical and Political Implications for Democracy

For decades, philosophers and ethicists have dealt with the well-known trolley problem – the moral dilemma in which one must decide whether to divert a runaway train to save several people at the cost of sacrificing one. Today, thanks to AI, this classical experiment takes on a new dimension. A recent study

(Fabre et al., 2024) has shown that moral advice provided by AI systems, used as moral advisors, significantly influences human decision-making: not only do people change their moral choices in such dilemmas, but they also alter the way they wish to be perceived by others. For instance, when AI systems put forward deontological arguments ("never kill directly," letting the train continue on its course and hit the group of people on the main track), participants became less inclined to choose the utilitarian option ("save the greatest number," by diverting the train and causing the death of a single person), partly to appear as morally irreproachable as possible in the eyes of others. These outcomes might give rise to significant political implications. First, they suggest that every public decision – from health policy to environmental governance – reproduces, on a collective scale, the same tension between the principle of utility and the principle of duty. Second, transposing the dilemma into the context of AI exposes a new form of technological governance: machines, which act in accordance with programmed moral parameters, embody ethical decisions that are themselves the result of prior political choices, made in advance (what counts as value, who deserves protection, and which outcomes should be optimized). Finally, the very figure of the AI-advisor challenges the notion of moral autonomy, which is deeply connected to political responsibility and democratic participation. Delegating moral decisions to a machine also means delegating part of one's sovereignty or civic agency – accepting that decisions about what is right or wrong may emerge from a non-human deliberation. Despite the obvious risks, the major idea is to view the involvement of AI-advisors in morally and politically relevant contexts as a potential instrument of democratization: a means to publicly confront, compare, and debate our ethical criteria. In this light, AI can become a political laboratory of responsibility – an opportunity to expand the spaces of cooperation and public reasoning, thereby making democratic action more self-aware, inclusive, and reflective.

Keywords: Ethical Governance; Democracy; Moral Decision-Making; Responsibility; Trolley Problem

Luca Tenneriello is PhD in Philosophy since 2020. His interests range from history of ethics to ethics of Artificial Intelligence. He has been a Visiting Researcher at the University of Kent, Canterbury (UK). He is currently a Post-Doc Fellow in Moral Philosophy at the Department of Computer, Control and Management Engineering, Sapienza University of Rome, working on a project on normative autonomy and AI. He has experience in counselling for AI ethical management (bias prevention, trustworthiness, ethical robustness against adversarial attacks). He is also an Executive Committee Member of the Italian Society for AI Ethics (SIpEIA).

Maran Ida Weeber
University of Amsterdam

False Memories in the Age of AI: The (In)Effectiveness of Warnings Against Hallucinated Misinformation

As generative AI systems like ChatGPT become increasingly integrated into everyday life, concerns have emerged about their potential to influence human cognition. This study investigated whether AI-generated misinformation (AI hallucinations) can induce false memories and whether explicit warnings about AI hallucinations can reduce this effect. A 2×2 between-subjects experiment was conducted, manipulating chatbot type (honest vs. rogue) and the presence of a warning (present vs. absent). Participants (N = 44) viewed CCTV footage of a robbery and were then interviewed by an AI chatbot. In the rogue condition, the chatbot subtly introduced five fabricated details. Memory accuracy was assessed through a subsequent

recall questionnaire. Results demonstrated that participants exposed to the rogue chatbot reported significantly more false memories for the manipulated details than those in the honest condition, confirming the powerful influence of AI as a suggestive agent. However, warnings about AI hallucinations had no significant effect on reducing false memory formation. These findings suggest that current warning designs are insufficient to counteract the persuasive effects of AI-generated misinformation and highlight the need for more robust safeguards in human-AI interaction contexts. As one of the first studies to test AI-generated misinformation in a false memory paradigm, these findings reveal the ethical challenges posed by AI's influence on human memory. They call for the development of robust, human-centered safeguards to ensure AI systems support and do not compromise users' cognitive autonomy.

Keywords: False memory; Generative AI; Chatbot misinformation; Suggestibility; Cognitive psychology

Maran Weeber is a recent graduate of the University of Amsterdam with a degree in Liberal Arts and Sciences, focusing on cognitive psychology and neuroscience. Her bachelor's thesis, which explored AI-generated misinformation and false memories, received a grade of 9.1 out of 10. She is currently a research assistant at the Free University of Amsterdam, working on studies involving people with intellectual disabilities. Her research interests lie in applied cognitive psychology and the ways psychological insights can inform policy and regulation. She is particularly interested in areas such as AI, where ethical and psychological challenges are closely intertwined.

SIpEIA

Società Italiana per l'Etica dell'Intelligenza Artificiale

con il patrocinio di

Associazione Italiana per l'Intelligenza Artificiale

#SIpEIA2026

www.sipeia.it