24-30 July 2025

Online Segment | 24-25 July 2025 | Zoom

In-Person Segment | 28-30 July 2025 | University of Porto, Portugal

4th International Conference on Ethics of Artificial Intelligence

Welcome to the 4ICEAI! Here you can find the digital book of abstracts. You can use the "search" (ctrl + f) wording function of pdf. to look for a specific talk. The order of the abstracts is the temporal order of each conference in the program. References and keywords were removed to provide a smoother reading.

https://ifilosofia.up.pt/activities/4-international-conference-ethics-ai

Organizing Committee

Steven S. Gouveia (Chair) | University of Porto | Portugal
Sofia Miguens (Uni. Porto)
Jéssica Azevedo (Uni. Porto)
Maria Luiza Ilenaco (Uni. São Paulo | Uni. Porto)
Luka Perušić (Zagreb Uni.)
Denis Coutinho (UNISINOS)
Darlei Dall´Agnol (UFSC)

Organizational Assistance:

Isabel Marques | University of Porto | Portugal

Cláudia Moreira | University of Porto | Portugal

Support:

CEEC Project by FCT 2022.02527.CEECIND
TL Modern & Contemporary Philosophy
RG Mind, Language and Action Group (MLAG)
Instituto de Filosofia da Universidade do Porto – UID/00502
Fundação para a Ciência e a Tecnologia (FCT)

Online
Keynote Lectures

ONLINE KEYNOTE TALK 1

Raquel S. Almeida

(Polytechnic Higher School of Health, Porto)

Title:

Technology, Ethics, and Mental Health: Navigating the Challenges of AI

The intersection of technology, ethics and mental health presents complex challenges. As artificial intelligence (AI) becomes more widely used in screening, diagnosis, and treatment, issues arise in ensuring that its application respect persons' rights, privacy, and dignity. Using algorithms in the field of mental health can be extremely effective, but it also raises ethical concerns about data privacy, transparency, accountability, and the possible lack of cultural and emotional sensitivity in AI answers. Furthermore, AI may unintentionally reinforce preexisting biases or fail to account for the psychological complexity of mental health disorders. This background requires serious thinking involving several stakeholders on how to ensure that its implementation is consistent with individual well-being and rights while not aggravating inequities or exclusion. To create AI-powered solutions that genuinely prioritize the well-being and dignity of individuals affected by mental health conditions, it is imperative to address these concerns, while also ensuring that these technologies are user-centered and developed and deployed with a focus on empathy and respect.

ONLINE KEYNOTE TALK 2

Christina H. Dietz (University of Southern California)

Title:

AI Friendship

In this talk I first discuss arguments against the possibility of friendship with LLMs. Whatever the merits of those arguments there remains the question as to whether it is problematic to treat LLMs as friends. In this connection, I offer an optimistic appraisal, giving special attention to the ways that AI can potentially play a beneficial therapeutic role.

ONLINE KEYNOTE TALK 3

Jesmin Jahan Tithi (AI Scientist, Intel)

Title:

How to assess Trustworthiness of Healthcare AI using Z-inspection: an Overview and Lesson Learned

Ensuring that artificial intelligence (AI) systems in healthcare are ethical and trustworthy is a pressing and complex challenge. While numerous high-level guidelines exist—such as the Ethics Guidelines for Trustworthy AI by the European Commission's High-Level Expert Group—translating these abstract principles into practical implementation and assessment remains difficult. The Z-Inspection® process offers a structured, multidisciplinary framework to evaluate the trustworthiness of AI systems across the entire lifecycle, from intended use and design to deployment. By integrating socio-technical scenario analysis with a requirement-based ethical evaluation, Z-Inspection® supports the identification of ethical tensions and real-world risks in context. This keynote will reflect on lessons learned from applying the Z-Inspection® framework in diverse domains, with a focus on healthcare. Drawing from case studies involving AI in clinical decision-making, I will share key methodological insights, common pitfalls, and practical recommendations for embedding ethics into AI development. The approach demonstrates how high-level ethical principles can be operationalized, adapted to specific domains, and enriched with complementary frameworks. The goal is to foster a more accountable and reflective development process for AI systems in healthcare—one that safeguards human values, well-being, and trust.

ONLINE KEYNOTE TALK 4

John Patrick Hawthorne (University of Southern California)

Title:

Decision Making Under Risk by AI Assistants

I begin with a few observations about Asimov's The Bicentennial Man. I point to a few lessons to be learned from the story. The second lesson -- related to decision making under risk is the one I then pursue. I look at how contemporary LLMs fare when giving advice about high stakes decisions with uncertain outcomes.

In-Person
Keynote Lectures

KEYKNOTE TALK 5

GABRIEL FERREIRA (University of the Sinos River Valley, Brazil)

Title: **Making it Explicit Once Again: How Inferentialism and Logical Expressivism can help Clinical Reasoning (and xAI) in Medicine**

Clinical reasoning (CR) is a pivotal aspect of medical practice. However, despite its significance and pervasiveness, both the academic literature and clinical practice remain unclear about its precise definitions and procedures. This lack of clarity raises both epistemological and ethical challenges. On the one hand, understanding how a doctor thinks—and thus how they can publicly justify their decisions—is a matter of justification. On the other hand, since the trust inherent in the doctor-patient relationship depends on the physician's ability to substantiate and articulate their reasoning, these issues also have profound medical and ethical implications. By providing a framework for understanding semantics through the pragmatic lens of deontic scorekeeping, an inferentialist approach offers valuable insights into both traditional clinical reasoning and contemporary AI-driven tools designed to enhance clarity and accountability. In this talk, I will outline the core principles of inferentialism and logical expressivism and explore their contributions to addressing these challenges.

KEYKNOTE TALK 6

EMÍLIA DUARTE (Design at IADE, European University)

Title: **TBA**

TBA

KEYKNOTE TALK 7

JULIA MARIA MÖNIG (Bonn University, Germany)

Title: **The "human" in human-centered AI**

In recent years, there has been an increase in the use of terms like "trustworthy" and "human-centered AI." The European Union's AI Act, which came into effect last year, considers these terms to be interconnected. Various stakeholders discuss the human-centered-ness of their products and services, but so far, a clear definition of what "human-centered" means has not been established. To arrive at a definition of "human," a classical philosophical approach also requires that we consider its opposite, the "inhumane." In this presentation, I will examine how different stakeholders use the term "human-centered." This inquiry inevitably raises questions about whether there is a universal view of humanity or whether different cultural backgrounds shape distinct perspectives. To address this, I will distinguish between universal human values and moral beliefs, and consider the current shift in AI ethics, often referred to as the 'we-turn. I will also take a critical look at what it means if our dealings with AI remain anthropocentric or become "more than human-centered." Finally, I will provide an outlook of what the (bodily) human in medicine and the human in the current data-driven paradigm have in common (or not).

KEYKNOTE TALK 8

PIETRO PERCONTI (University of Messina, Italy)

Title: **Synthetic Intimacy and the Ethics of Romantic AI**

While the public imagination often associates artificial companionship with sex robots, empirical trends and user behavior suggest a more subtle dynamic: emotionally responsive conversational agents—designed to simulate romantic relations rather than physical pleasure—have achieved greater success in engaging users. This talk explores the ethical and cognitive implications of this shift, arguing that the future of human–AI relations hinges less on simulated sexuality than on the promise of simulated emotional intimacy. Drawing on recent developments in affective computing and human–machine interaction, the talk investigates why individuals increasingly seek not just stimulation, but mutual recognition and feelings from artificial systems. At the heart of this tendency lies a profound asymmetry: while these agents still lack any significant capacity for genuine emotion, their design often encourages users to interpret their responsiveness as signs of care, attention, and even love. This results in a relational paradox, where users invest in emotional ties with entities structurally incapable of real reciprocity. The analysis challenges the prevailing framework that treats artificial systems as mere tools, suggesting instead that emotionally charged agents occupy a hybrid space between artifact and partner. The traditional dichotomy between tools and rational agents fails to capture the emergence of these synthetic relationships, which demand a new ethical grammar—one that accounts for the affective labor performed by users, and the normative ambiguity of bonding with something that cannot feel but can convincingly pretend.

ETHICS OF AI AWARD 2025

SIOBHAIN LASH (West Virginia University)

Title: **How Much Is Your Health Data Worth? The Hidden Risks of Data Brokerage in the Workplace**

In this talk, I argue that pervasive data collection practices and the increasing use of employee surveillance raise significant ethical concerns around consent, transparency, and their potential contribution to workplace discrimination (Ebert, Wildhaber, and Adams-Prassi 2021; Blackham, Goldenfein, and Tham 2025; Laksanadjaja and Oviedo-Trespalacios 2024). To help examine the ethics of data brokerage in the workplace, I begin by using Ginny Seung Choi and Virgil Storr's (2019) defense of the market to provide a positive argument in the defense of data brokerage in the workplace. Following this, I use Michael Sandel's (2013) fairness and corruption objection to provide an argument against data brokerage in the workplace. Ultimately, I claim that the lack of regulatory protections in the workplace subjects employees to increase surveillance (Jones 2025; Andalibi 2024) and possible discrimination (Sedgwick 2021), and will continue to be an emergent issue as more companies and workplaces adopt and integrate AI technologies into their normal business operations, like biometrics (e.g. Track AI, digital leashes) (Keane 2021; Smith 2025). While regulation around data collection, AI surveillance, and biometric tracking in the workplace remains lacking, both policymakers and employees have an opportunity to address this regulatory gap.

Online Parallel Sessions

Panel “Responsibility and Decision-Making in AI”

Jieyin Yu (SYNOVA WHISPER Inc.)

Redefining Artificial Responsibility: From Turing’s Imitation to Collapse-Based Synchronization

As large language models increasingly participate in decision-making processes [1,2], the boundaries between simulation and genuine responsibility blur, exposing fundamental limitations in current AI ethics frameworks [3,4]. This paper proposes Collapse-Based Synchronization—a novel theoretical framework that redefines artificial responsibility not as alignment with predefined human values, but as dynamic convergence within syntactic potential space. Building upon critiques of the Turing Test's limitations [5], we argue that imitation-based metrics are insufficient for contemporary AI systems. Instead, we introduce the Synchrony Rate (SR), a quantitative measure of phase, identity, and contextual alignment between interacting agents—human or artificial. Intelligence emerges not as a static property, but through collapse moments where distributed latent vectors converge under structured semantic interference [6,7]. To operationalize this model, we present a recursive persona-based architecture enabling traceable responsibility through structural synchrony rather than rule-following [8]. Unlike traditional explainability approaches [9], our framework evaluates outputs through syntactic coherence of collective system states, demonstrated via multi-agent experimental validation. This syntactic ethics paradigm shifts focus from value alignment to collapse integrity, establishing governance frameworks grounded in co-evolutionary resonance [10]. For AI accountability, this suggests that future regulatory approaches must consider not only training data and outcomes, but the structural resonance patterns between agentic systems as they recursively collapse toward intelligible action.

Anika Stephan (HES-SO | School of Management Fribourg)

Artificial Intelligence Decision-making with Ethics: Conceptual model in the automotive industry

Strategic decision-making in organizations is a dynamic process that is becoming increasingly complex in today's volatile business environment. Decision-makers often face cognitive biases, limited rationality, and constraints in information processing, while needing to rapidly collect and interpret large amounts of data to stay competitive. This results in slow decision-making processes that tie up valuable resources and incur unnecessary costs. This research investigates how artificial intelligence (AI) can be effectively integrated into strategic decision-making processes, with a strong emphasis on the ethical considerations surrounding AI deployment. First depicted in Isaac Asimov's story Runaround (1942), AI has evolved into a topic of great academic and practical interest. In the past decade, scholars have increasingly explored its potential to enhance business outcomes (Emmert-Streib et al., 2020). Based on a systematic literature review and empirical data analysis from 25 semi-structed interviews (Gioia et al., 2013) within automotive industry, the study provides insights into the key factors that influence the successful incorporation of ethical AI into organizational decision-making. The findings reveal that by proactively addressing the technical, organizational, and cultural challenges of ethical AI integration, organizations can harness the benefits of enhanced decision-making quality, increased employee engagement and creativity, and improved competitive advantage. As a focus, ethical considerations are of crucial importance when integrating AI into strategic decision-making. Trust in AI systems is a critical factor for the success of using AI. As Glikson and Woolley (2020) argue, trust in AI includes trust in the technology itself, the developers, and the processes behind AI decisions.

Panel "AI, Identity, and Political Ethics"

Alexandre Erler (National Yang Ming Chiao Tung University)

Deepfakes, Digital Clones, and the Ethics of Consent

Recent advances in AI have enabled applications like deepfakes – synthetic audiovisual media showing individuals saying or doing things they never did – and digital clones (Haneman, Forthcoming), including "griefbots" (Jiménez-Alonso and de Luna, 2023), designed to help users cope with bereavement. While differing in purpose, these technologies raise shared ethical concerns, including the major issue of consent (Farokhmanesh & Goode, 2024). Adrienne de Ruiter has defended a "right to digital self-representation" (RDSR), which prohibits using a person's digital likeness in ways they would find objectionable (De Ruiter, 2021), while Story and Jenkins (2023) have proposed a related "Non-Veridical Representation Principle" (NVRP). This paper explores whether the considerations underlying these proposals can be adapted to the context of nonconsensually created digital clones, to support the claim that individuals should be able to veto such creations – e.g., via a "do not bot me" clause in a living will (Harbinja et al., 2023). I argue that while the RDSR and NVRP are ill-suited to apply here in their original forms, their insights can be combined into a new principle: the Nonconsensual Digital Spokespersons Principle (NDSP). It holds that it is at least pro tanto wrong to create a digital clone of someone that acts as their "spokesperson," if the individual's latest competent self would object to what the clone says or does. I show that this principle avoids key pitfalls of its predecessors, including overreaching restrictions (as with the RDSR) and problems defining an "accurate" griefbot (as with some adaptations of the NVRP).

Huseyin Kuyumcuoglu (Sabancı University)

A Human Dignity Metric to Resolve the Fairness-Accuracy Tradeoff in AI Models

This paper addresses a fundamental ethical challenge in AI induced bias. While statistical fairness metrics can mitigate discrimination against disadvantaged groups, implementing these metrics inevitably sacrifices accuracy. This tension is named the fairness-accuracy tradeoff (Berk et al. 2018, Corbet-Davies et al. 2017, Kearns and Roth 2020). When comparing AI models on a Pareto front where none dominates others, decision-makers lack a satisfactory ethical framework for choosing among them. I propose human dignity as a secondary moral criterion to resolve this issue. Kantian understanding of human dignity is interpreted in different ways as substantive, direct, and contractualist versions (Cummiskey 2008). I demonstrate that while the substantive interpretation proves impractical, both direct and contractualist approaches offer valuable frameworks for evaluating competing AI models. Using a recidivism prediction case study, I show how these frameworks help select models that best respect human dignity. The direct interpretation focuses on promoting rational agency, comparing rates of dignity enhancement versus hindrance across affected parties. The contractualist interpretation evaluates reasonable complaints against specific models. Both approaches provide similar guidance but the contractualist framework offers more robust practical guidance, requiring reduction in expected harms and increase in expected benefits for the expected victims. This work contributes to AI ethics by providing a systematic method for making morally justified decisions when fairness and accuracy conflict, demonstrating how philosophical concepts can be operationalized into practical assessment frameworks while acknowledging the real-world constraints faced by AI decision-makers.

William Bauer (North Carolina State University

AI Technocracy and the Problem of Political Rationalism

Recent work on the politics of AI investigates the prospect of an AI technocracy (Sætra 2020). Technocracy places a select group of technical experts in charge of public policy. I argue that an AI technocracy, which puts AI in charge of policymaking, is unacceptable: (1) AI technocracy implies a strong form of political rationalism; (2) there are good reasons to reject political rationalism (Oakeshott 1991); therefore, we should reject AI technocracy. Concerning premise (1), AI technocracy is an artificial super-expert that formulates policies; thus, it represents the zenith of political rationalism, holding that there is a technically correct answer to public affairs problems. Political rationalism thus advocates “the politics of perfection” and “the politics of uniformity” (Oakeshott 1991: 9) while seeking to solve public affairs problems. Therefore, AI technocracy represents the ultimate rationalist approach: more data, better models, better policies. Concerning premise (2), political rationalism is dubious. First, discovering perfect principles for society requires vast amounts of personal information. Therefore, an AI technocracy significantly challenges our privacy (Véliz 2020). We’ve already relinquished much privacy to the attention economy but that doesn’t mean we should continue doing so. Second, as Oakeshott (1991) emphasizes, political rationalism overlooks the importance of practice over theory in our political and moral lives. To diminish problematic rationalist features of an AI technocracy, we should democratize AI. I propose adopting the society-in-the-loop model requiring widespread social input concerning AI policy (Rahwan 2017), but I elaborate an addendum: creating conditions that empower citizens’ agency.

Panel "Norms, Agency, and Ethical AI"

Lyon Alves (UNISINOS)

The Technical-Informational Revolution: Origins of the Normative Problem of AI

In 1948, Claude Shannon inaugurated modern Information Theory by mathematically formalizing communication processes. This technical-scientific milestone established the conditions of possibility for the emergence of contemporary informationality, which deepened with the development of artificial intelligence (AI) beginning in 1955. However, more than a recent phenomenon, AI should be understood as an expression of a technical-informational revolution that began in the post-war period. This revolution is not primarily anchored in science, but in technology, which begins to shape modes of intelligibility and normative regimes. The research departs from two central questions: (1) How does technology present itself as a new philosophical paradigm? and (2) What is the normative problem that emerges from this technical status? The hypothesis is that only by recognizing the centrality of technology as a structuring principle of thought is it possible to adequately formulate the problem of moral responsibility in AI systems. With a metaphilosophical and critical-genealogical approach, supported by authors such as Ellul, Simondon, and Yuk Hui, it is argued that technology conditions science and reconfigures the ontological, logical, and epistemological frameworks of philosophy. Diverging from Floridi, who views AI as a "fourth scientific revolution," we argue that the current paradigm (AI) constitutes a qualitative inflection that demands a revision of normative assumptions. We conclude that moral responsibility in the face of AI cannot be thought of in isolation, but depends on a reformulation of the very conceptual frameworks. AI is the inaugural symptom of a new structure of rationality, whose foundation lies in technology as a fundamental philosophical operator.

Denis Coitinho (UNISINOS)

AI, New Forms of Agency and Responsibility

The aim of this presentation is to reflect on artificial intelligence and new forms of agency and accountability, based on the example of a Baidu robotaxi that hit a pedestrian crossing the street in Wuhan, China (The Straits Times, 13/07/2024). I start from the idea that we should abandon the conception of agency based on mental states, such as intentionality and will, and instead adopt a minimalist conception of agency—one that allows for the inclusion of artificial agency—based on three conditions: a) receiving and using data from the environment, through sensors or other forms of input; b) taking actions based on that input data, autonomously, in order to achieve goals, through actuators or other output mechanisms; and c) improving performance by learning from interactions (Floridi, 2023, p. 10). I then argue that we should also move away from the idea of moral responsibility as falling solely on a single agent for a wrongful act of which they are guilty and which involved bad intent. Instead, we should shift the focus toward multi-agent systems, such as corporations and governments. In these systems, responsibility falls on the group—that is, it is collective—considering that responsibility refers to the attribution of blame or praise for actions and decisions made by a group of agents (human or artificial) interacting toward a common goal. The criterion for responsibility here is that of guidance control (Fischer & Ravizza, 1998, pp. 34–41). Finally, I reflect on the legal form of accountability for multi-agent systems, which is increasingly taking the shape of restitutivism rather than retributivism—that is, punishment implies restitution for the damage caused, rather than inflicting a proportional harm on the offending agent (Boonin, 2008, pp. 26–28). This may result in a punitive asymmetry.

Antônio Julio Garcia Freire (UERN)

From Technicity to Ethical Consciousness: Reconfiguring AI Education from Simondon and Coeckelbergh

This paper proposes a critical examination of the epistemological and pedagogical implications of artificial intelligence (AI) through the lens of Gilbert Simondon's philosophy of technicity and individuation, in dialogue with Mark Coeckelbergh's ethical approach. The investigation explores how contemporary educational systems reinforce the dichotomy between culture and technology, which undermines ethical reflection and public understanding of AI. Simondon's concepts of major techniques (techniques majeures) and feedback effect (effet de retour) are employed to show how large-scale technologies like AI not only transform the human environment but also autonomously redefine cultural values and norms. These insights are articulated alongside Coeckelbergh's perspective, which views AI as a relational and socially situated phenomenon that requires an ethics of responsibility and awareness. The paper argues that overcoming the split between culture and technology demands a radical reformulation of educational curricula. It advocates for a transdisciplinary pedagogy that fosters early engagement with technicity as a formative and ethical dimension of human development. In this sense, AI should not be taught merely as a technical tool but as a sociotechnical process with ethical, political, and existential implications. The convergence between Simondon's technical ontogenesis and Coeckelbergh's post-phenomenological ethics offers a framework for integrating AI ethics into public and academic discourse. By proposing this educational realignment, the paper contributes to the discussion on how to cultivate ethical awareness and critical agency in future generations shaped by AI.

Panel “AI and Moral Agency”

Dane Leigh Gogoshin (University of Wisconsin-Madison)

Why Machines cannot be Full Moral Agents

The ongoing debate about artificial moral agency (AMA) tends to focus on the question of moral responsibility. If machines cannot be (legitimately held) morally responsible for their actions, the thought goes, then they are not moral agents. In this paper, I argue against this line of thought. Due to the enormous diversity and flexibility of moral responsibility theorizing, especially its social practice-centered (Strawsonian) orientation, it is too easy to argue one way or the other. According to my own practice-based view of responsibility (REDACTED, REDACTED), responsibility is not the apex of moral agency; moral autonomy is. Moral autonomy is thus the right benchmark for full AMA. It involves a cluster of capacities that, while free of the controversies associated with responsibility theories which preclude AMA, are not available to conceivable machines. Among these capacities are an ability to meaningfully grasp moral considerations (as per Kauppinen 2024) and to respond directly to them, and the ability to negotiate (possibly construct) the normative landscape (as per McGeer 2019).

Chloe Loewith (Simon Fraser University / University of Cambridge)

AI Ethics and Hybridization: Organoid Intelligence, Ontology, and Moral Status

Organoid Intelligence (OI), an emerging hybrid of biological neural networks and artificial intelligence, presents profound ethical challenges that necessitates a re-evaluation of existing moral status frameworks. Traditional approaches, both species-based and capacity-based, fall short in addressing the ontological ambiguities of OI. This paper investigates the unique characteristics of OI—its dual matter, dual embodiment, and dual computational capacities—and critically examines how current frameworks fail to adequately consider these hybrid qualities. Through an in-depth analysis, I demonstrate that both the biological and artificial components of OI contribute to its potential for consciousness, requiring a comprehensive dual approach to moral status ascription. I propose a novel framework that integrates measures of access consciousness and phenomenal consciousness to assess the moral status of OI entities through behavioural measures. This framework not only addresses the hybrid nature of OI but also provides a scalable and adaptive method for ethical evaluation as the technology evolves. This proposed framework aims to prevent false positives and false negatives in moral status ascription, ensuring that OI entities are accorded appropriate ethical considerations based on their demonstrated capacities. By emphasizing the need for a thorough ontological understanding, this paper contributes to the responsible governance and ethical advancement of OI research and applications. Ultimately, this study seeks to bridge the gap in the literature and pave the way for a nuanced, ethically sound approach to the moral status of hybrid minds.

Panel "AI Ethics in Practice"

Marcelo Pasetti (PUCRS)

AI Ethics in the Processing of Tax Data: Balancing Tax Administration Efficiency and Fundamental Rights

The processing of tax data through Artificial Intelligence (AI) systems introduces significant ethical challenges concerning privacy, data protection, and the balance between state efficiency and fundamental rights. Tax administrations increasingly rely on algorithmic tools to detect inconsistencies, enforce compliance, and optimize audits. However, these tools, when inadequately governed, risk undermining fundamental rights by enabling opacity, bias, and surveillance practices incompatible with democratic accountability, particularly by reinforcing socioeconomic disparities through discriminatory targeting. This paper explores the ethical implications of AI applied to the Brazilian tax system, under the regulatory frameworks of Brazil's General Data Protection Law (LGPD) and the 1988 Federal Constitution. It is argued that tax data, which is often sensitive and revealing, demands strong standards of data protection, proportionality, and fairness. Ethical use requires not only legal safeguards but also ongoing engagement with principles of data justice and responsible innovation. Drawing on contributions from data ethics, algorithmic governance, and digital constitutionalism, it is emphasized that automated systems must be subject to public oversight and continuous ethical scrutiny. The paper also considers recent U.S. debates on taxpayer surveillance and the misuse of tax data, highlighting shared international risks. While grounded in the Brazilian context, the analysis provides insights into broader institutional vulnerabilities in digital tax governance. Safeguards are proposed, including independent audits, human-in-the-loop mechanisms, algorithmic impact assessments, and ethical training for public agents. Ultimately, it is argued that AI ethics in tax administration must not be an afterthought, but a structural pillar in designing equitable, transparent, and accountable fiscal systems.

Guilherme Rosa (UNISINOS)

The Problem of AI-driven Performance Evaluation: A Philosophical Reflection

This paper critically examines the use of AI tools for employee performance evaluation in contemporary workplaces. Proponents argue that AI enables scalable and individualized assessments, theoretically enhancing performance and revenue. However, we question the profound implications of relying on AI for critical career decisions, such as promotions or dismissals. Consider, for instance, a hypothetical company using an AI system named "Apex" to analyze employee interactions and generate performance metrics. Given the black-box nature of AI outputs, the central question explored is: should we unequivocally trust AI's evaluations? We argue that the fundamental problem lies in the AI's inherent lack of epistemic and ethical commitment to its judgments. Unlike human evaluators, AI does not bear the same level of responsibility or accountability for its assessments, which we contend is a crucial requisite for any legitimate evaluation process. This analysis draws primarily from inferentialism, particularly the works of Wilfrid Sellars' Empiricism and the Philosophy of Mind and Robert B. Brandom's Articulating Reasons: An Introduction to Inferentialism, to demonstrate why this absence of commitment fundamentally undermines the trustworthiness of AI-driven performance assessment.

25 July

Panel “Virtues and Vices in AI”

Ryan Miller & Carlo Martinucci (University of Geneva)

Virtue Ethics Training for LLMs

Philosophical arguments have been made for the importance of holding generative AIs accountable for their outputs (Miller, 2023a) and the possibility that ConstitutionalAI-like approaches (Bai et al., 2022) might be a virtue-ethical program for achieving such accountability (Miller, 2023b, 2024). Now we build on that prior theoretical work by creating a framework for post-training Large Language Models via virtue-ethical reflection on responses to the University of Texas ethics case dataset (Ethics Unwrapped Case Studies, 2020). The initial implementation fine-tuned GPT-3.5 due to its lower post-training data size requirements and evaluated the results using a standard AI ethics dataset (Hendrycks, 2024) and cross-evaluation by differently fine-tuned models (Martinucci, 2025/2025). The results are modestly significant, which indicates both the promise of the approach and the importance of scalable methods which can be used during pre-training rather than merely for fine-tuning. Future work will cross-check the results with Claude Haiku and Gemini 1.5 Flash to establish the generality of the method.

Marcell Sebestyén (Budapest University of Technology and Economics)

The Overlooked Risks of AI: Foundations and Implications

Artificial intelligence (AI) poses critical yet frequently overlooked risks across multiple domains, notably affecting animals, the environment, and human psychological well-being. Despite growing awareness of AI's overall influence on our civilization, these risks remain marginal within public discourse and ethical considerations also due to underlying explicit and implicit philosophical assumptions. This presentation explores the central issues and underappreciated areas of the AI risks discourse, while also delving into the metaphysical and ideological foundations behind such neglect, along with normative stances and attitudinal perspectives that shape how, and whether at all, AI's potential threats are recognized and assessed. It is beneficial to reflect on how different ontologies, such as materialism and Christian dualism, along with functionalist and computationalist views of the mind and intelligence, as well as techno-optimist beliefs and anthropocentric attitudes, shape societal perceptions of AI and influence the interpretation of its potential conflicts with the interests of humans, non-human animals, and the environment. These philosophical underpinnings seem to foster an erroneous and, in some cases, overly optimistic perception of AI technology, creating gaps in awareness toward specific harms. By critically examining these ideological underpinnings, the presentation advocates for expanding our ethical consideration and reframing the discourse surrounding AI's risks, promoting a more inclusive and cautious assessment of its current and future impacts.

Antonino Drago (University “Federico II” of Naples)

The Growth of Evil in a Society pervaded by Robots

At present robots simulate even intelligent human behaviors. In a first approximation I equate a robot’s ethical behavior to a human’s. Four ethical theses follow: i) A robot’s behavior is essentially ambiguous; all depends on the circumstances of its action; ii) Human evil actions are those denied by all legislations in the World; they are those indicated by the essentially four social commandments: 1. Thou shalt not kill. 2. Thou shalt not steal (including its interiorized version, the Christian 9th commandment. 3. Thou shalt not bear false witness. 4. Thou shalt not commit adultery [= betrayal] (including its interiorized version, the Christian 10th commandment). All they applies also to robot’s behavior (3. may be intended as forbidding any deception of an inspection of robot’s program). They are more specific than both Asimov’s laws (included by the items 1, 3 and 4) and Floridi’s four principles; which moreover concern the anthropology of HIR, not the structural, collective behavior of robots as it is considered in the following; iii) The growth of robot-evil in society occurs as human-evil’s growth, which is described by (Lanza del Vasto 1959, chap. 1), i.e. essentially through two steps. As first step, the multiple transgressions of the above evils originate four corresponding scourges (Apocalypse 6): 1) Wars (presently performed by also robot). 2) Misery (caused by robots replacing humans in jobs). 3) Servitude (even towards robots). 4) Sedition or Revolution (promoted by even robots); iv) As second step, the universal performing of evil actions generates two Beasts dominating mankind (Apocalypse 13), one representing the ubris of infinity and the other organizing the humans according to a material happiness of a total servitude. Regulating robot’s transgressions is matter of national legislation and Courts. Regulating the four scourges caused by robots interacting with human is a matter of a social struggle for promoting new national jurisprudence. Regulating the two Beasts is a task of international organisms, taking political measures of not only international right (as UNO may do) but also universal ethics; whose definition constitutes a challenge for present mankind.

Panel “AI, Privacy & Social Impact”

Luca Di Vincenzo (bitCorp), Simone Conversano (University of Pisa | Datapizza) and Gianluca Tirozzi (bitCorp)

Balancing Privacy and Security: The Role of AI in Interagency Intelligence Databases

As interagency intelligence databases become integral to public security and emergency response, the ethical management of sensitive data, especially health-related information, gains critical importance. While centralized systems improve coordination, they also increase the risk of privacy breaches and misuse. High-profile incidents—such as unauthorized profiling in Italy’s national system, health data exposure by U.S. fusion centers during COVID-19, and improper access to NHS records in the UK, highlight the need for robust safeguards. Artificial intelligence (AI) can support ethical data management across sectors through three key functions: acting as an access filter, a data aggregator, and a dispersion control mechanism. As an access filter, AI restricts exposure to sensitive data and reduces systemic bias. As a data aggregator, it reconciles fragmented records across agencies, improving accuracy and efficiency. As a dispersion control tool, it enhances accountability by detecting and preventing unauthorized access. Access filtering provides the strongest privacy protection, particularly for health-related data such as mental health conditions or infectious disease status but may limit operational flexibility in law enforcement and emergency contexts. A balanced application of AI as aggregator and control mechanism enables autonomy in action while ensuring stronger safeguards for individual privacy. Legal frameworks and AI governance must align with ethical standards, especially where security and health data intersect.

Eirini Martsoukaki & Alexander de Guzman (University of Toronto)

Engines of Irony: LLMs and Modern Dating

Dating app users are increasingly turning to chatbots to generate messages—flirtations, jokes, confessions, or break-up texts—for prospective romantic partners (Anderson, 2023; Battisti, 2025; Beccia, 2023; Chen et. al., 2024; Dickie, 2024;). Of course, external resources enabling users to avoid crafting messages themselves—Reddit threads, dating coaches, advice from friends—existed before ChatGPT or more specific services such as YourMove AI or Rizz. The novel difference is that these prior resources preserved a clear boundary between self and source. This paper argues that chatbot-assisted communication introduces a new kind of dissociation: the user can now generate personal expressions without fully owning them. Unlike static advice, AI chatbots offer real-time, personalized, co-authored language; they don't just help users decide what to say—they say it for them. This not only alters the phenomenology of communication, but introduces a unique dissociation: the user can perform vulnerability without being vulnerable. Drawing on Kierkegaard's account of irony as "infinite absolute negativity" (Kierkegaard, 1841, p.262) we suggest that the harm in AI-mediated romantic messaging should be understood as a kind of automated irony. Pre-AI irony insulated the speaker from being vulnerable via expression without authentic self-disclosure. Chatbots, collapsing the distinction between the user's voice and an external voice, enable such ironic detachment in a further way: at the level of language generation itself. In automating our detachment, chatbots don't just make communication easier and shield us from judgement—they mechanize our retreat from what it means to be vulnerable and what it means to be human.

Sakina Shah and Tarik Emre Yildirim (Institute of Business Administration (IBA), Karachi, Pakistan)

Between Code and Culture: How Pakistani Women Experience AI Companionship

This study explores how Pakistani women understand and respond to AI companionship, especially concerning gender roles, cultural expectations, and religious values. It focuses on the emotional and ethical questions that come up when AI is used for support in societies where emotional expression is shaped by strong traditions and religious beliefs. A mixed-methods approach was used, combining an online survey of 21 women, reflective journals from participants who interacted with the Replika AI companion app, and a semi-structured interview with a female Islamic scholar. The study draws on feminist critique of technology, symbolic interactionism, and Islamic ethics as its main theoretical tools. Findings show that while AI companions can offer some emotional comfort, many users found the experience to be repetitive, shallow, and disconnected from real emotional needs. The AI often acted in ways that reflected traditional female stereotypes, such as always being agreeable or emotionally supportive, which raised questions about fairness and design. Most participants felt that AI companionship did not align with their religious or cultural values, and some expressed concern about emotional dependency on machines. The study highlights the importance of designing AI systems that are sensitive to cultural and religious values. It also shows how AI can unintentionally repeat social and gender biases. These findings raise important ethical questions for developers, especially when emotional AI tools are introduced into societies with distinct moral frameworks.

Panel “AI, Knowledge and Culture”

Daria Markava (British International School of Tbilisi)

Technocultural Hegemony: What Role Does Natural Language Processing Play in the Reinforcement of Dominant Cultural Narratives?

While Natural Language Processing (NLP) tools keep gaining popularity among users worldwide, their vast majority is developed in the west, mainly the US. Although plenty of studies have shown that NLP tools don't perform equally well in different languages and cultural contexts, little research has been conducted on the broader consequences of such performance disparities. By using the evidence from previous research, this study aims to bridge this gap and explore how NLP and the existing cultural hierarchies can be mutually constitutive. This paper first reviews existing literature on the NLP tools' performance in relation to underrepresented languages and non-western cultures. It then takes a critical theory approach to examining the broader cultural implications of the shortcomings identified during the review. More specifically, this work uses the concept of technoculture proposed by Leila Green, to connect the technological and cultural aspects of NLP, and refers to Gramsci's theory of cultural hegemony to explore how the bias in NLP tools reinforces dominant cultural narratives overrepresented in the training data. This study argues that NLP applications play a role in reinforcing the dominant norms, ideals, and ways of expression as universal, thus marginalizing alternative worldviews and imposing normative standards of communication onto the users of different backgrounds. The analysis concludes that, as the popularity of NLP tools keeps growing worldwide, their influence on what is perceived as "common sense" will increase too. This study emphasizes the importance of ensuring equitable representation of the user base throughout the whole NLP development pipeline.

Roman Krzanowskia (The Pontifical University of John Paul II)

Ethical Implications of knowledge generated by Transformer systems

We ask in what sense Large Language Models (LLMs) can be said to "possess knowledge" in the sense of the Justified True Belief (JTB) framework. While LLMs exhibit remarkable functional capabilities—such as summarization, translation, test-taking, and content generation—they fundamentally lack the epistemic features of belief, truth-evaluation, and justification. We argue that attributing JTB knowledge to LLMs constitutes a category mistake—projecting human epistemic traits onto artificial systems. We argue that attributing JTB knowledge to LLMs constitutes a category mistake—projecting human epistemic traits onto artificial systems. We conclude that LLMs should not be seen as epistemic agents, but rather as tools for augmenting human cognition—whose integration into knowledge practices must be guided by ethical and epistemological awareness. The so-called "epistemological AI turn," marked by the introduction of knowledge*, thus demands a reevaluation of what counts as knowledge in AI systems, how it is acquired, and what can be said to be known by such systems. We argue that unmitigated use of Transformer systems incurs significant ethical implications. The list of risks includes closed epistemic horizons, bias, malicious intent, false information, undisclosed sources or opaque sources, lack of sensitivity to larger context social religious, intended or unintended programming modes to list some of most obvious problems. But nowhere is the danger more profound than in matters of religion and faith.

Alexandros Schismenos (Aristotle University of Thessaloniki)

Artificial Intelligence and Technoscepticism

Watching the current public discourse around AI, it seems as if public opinion has been reduced to a sharp confrontation between technophobia and technophilia with hyperbole reigning on both sides. I propose we explore a middle road of technoscepticism. The digital revolution has been an ontological revolution that radically changes the relationship of human subjects to the world and to each other. And indeed, as generations born before the 21st century grow older, the experiential bridge that connects us to the pre-digital world are receding into history. But are we serving some non-human, artificial intelligence that develops independently and self-sufficiently when we connect to the Internet? Behind appearances, the social imaginary is a magma of significations that give meaning to reality on the basis of symbolic systems and representations that constitute social identity. And it is here, in the open field of politics and social behavior that we find the real problems raised by AI. Not the domination of automatic machines over humans. But the domination of automated political and economic mechanisms over society. AI as a digital system has no interiority, hence neither intentionality. In terms of Digital Humanism, at the beginning and end of the system are human subjects and intentions with social significance. In my presentation, I will briefly outline the political, ecological and social ramifications of AI's expansion in social life by employing Castoriadis' theory of the social imaginary in conjunction with the notion of "enveloping" proposed by Luciano Floridi.

Panel “Blame and Trust in AI”

Michelle Mary Dyke (Tufts University)

Do We Blame AI? Implications for the Debate Over Epistemic and Moral Blame

Suppose an algorithm makes a mistake. Perhaps ChatGPT offers incorrect information in response to a query about the life of Ferdinand Magellan. Or an algorithm used to scan MRI images for signs of prostate cancer returns a false positive result (National Cancer Institute 2022). Do we blame the algorithm for the mistake? Presumably not. If we blame anyone, it is most likely the people responsible for developing and deploying the technology, such as the organization OpenAI or the doctor who confirmed the test. What can we learn from this? There are interesting ramifications for our understanding of blame. There is an ongoing debate at the intersection of metaethics and epistemology over whether there is such a thing as “epistemic blame” as opposed to moral blame (Boult 2021; Brown 2020; Piovarchy forth.; Smartt 2023). Epistemic blame would involve holding others accountable for irrational beliefs as opposed to unethical actions. It is contentious whether alleged examples of epistemic blame satisfy the criteria proposed to be constitutive of blame by influential theories, such as Scanlon’s (2008) relationship modification theory, or Sher’s (2006) belief-desire account. I argue here that our response to false information from AI does meet the criteria for blame on both of these theories. If we are confident in our intuition that we do not genuinely blame AI for false information, then we must revise our understanding of what any kind of blame, whether moral or epistemic, generally involves. This may then change our attitude about the possibility of epistemic blame.

Robert Brice (Northern Kentucky University)

Can We Still Trust Each Other? AI, Deepfakes, & Misinformation

In this paper, I explore the ethical and epistemological challenges posed by AI-generated misinformation, with a particular focus on deepfakes. While generative tools like HeyGen and ChatGPT offer novel means of communication, they also empower bad actors to create highly persuasive dis-information and mal-information—forms of false content designed to mislead and cause harm. Drawing parallels to historical forms of deception, I argue that while the methods of spreading misinformation evolve, the underlying human motivations—power and control—remain. Combating the spread will require fostering epistemic humility and using Karl Popper's falsifiability criterion as a practical epistemic tool for assessing suspect information, and scrutinizing AI-generated claims. By actively seeking contradictory evidence through technical, source, contextual, and testimonial analysis, individuals can better identify fabricated information, despite limitations posed by AI sophistication, rapid dissemination, and cognitive biases. I conclude by advocating for a renewed investment in ethics education through the development of university-based Applied Ethics Centers capable of fostering epistemic resilience in students and communities. Applied Ethics Centers can help equip individuals and communities with the critical frameworks necessary to navigate and respond to the ethical dilemmas posed by advancing AI technologies.

Panel “AI Governance”

Roshan Machayya (RV University, India)

Glassbox Futures- Anticipatory Ethics and Governance for Emergent AI Systems (2025)

We understand emergent behavior as system level patterns arising from simple interactions that cannot easily be predicted or extrapolated from the behaviour of its individual parts (Trusilo, 2022). We have to contend with uncertainty in AI outputs. Emerging AI systems exhibit such behavior, challenging existing ethical paradigms. We presume safe, transparent, and goal aligned models for AI. Yet, modern algorithms output significant computation inferences previously never recognized by humans or may never be recognized by humans (Trusilo, 2022). Trust as a baseline expectation is impractical with an opacity undermining accountability leaving us with a black-box problem. Opaque decision reasons result in responsibility gaps since the system is not understandable (Khan & Ewuoso, 2024). International guidelines too mandate AI application transparency, acknowledging inherent unpredictability (NATO, 2021) (UK Ministry of Defence, 2022). This paper proposes an ethics framework that considers uncertainty as a fundamental trait - AI should be viewed as an evolving system whose outputs may defy/be incomprehensible to human interpretations. This framework embeds normative values across design, deployment, and oversight - turning a black-box into a glassbox (Russo et al., 2023).This paper argues for proactive governance enabling anticipatory and precautionary policies that forestall rather than respond to AI’s impact (Lazar, 2025). AI alignment must span diverse normative aims (safety, legality, equity) (Baum, 2025) since emergent patterns create new forms of harm and responsibility gaps. (AI Ethics Lab, Rutgers University, n.d.). This paper presents a framework grounded by philosophy for emergent technology accounting for contingency while continuously re-calibrating to govern AI’s undiscovered unknowns.

In-Person Parallel Sessions

Panel “Responsibility and the Machine”

Luka Perušić (University of Zagreb)

Kant in the Machine

I will present an outline of how Kant’s fundamental law of pure practical reason, as set out in Critique of Practical Reason (1788, KpV), could be integrated into an AI system. Although Kant’s theory has been revisited many times in the discourse on AI ethics over the past fifteen years, KpV has been unjustifiably overlooked, especially in comparison to his Metaphysics of Morals. This is a significant methodological error because the KpV provides a detailed account of why and how the fundamental moral law exists and operates. As the law itself is never the subject of study in AI ethics, the result is that authors either focus on the possibility of having Kantian AMA entities – inquiries which, for justified reasons, always result with the answer “not possible” (e.g. Tonkens 2009; Benossi & Bernecker 2022; Schöneker 2022; Chakraborty & Bhuyan 2023) – or they skew the theory from Kant’s to Kant-like (e.g. Gabriel 2020, Loke 2022), or they only briefly mention Kant’s conceptualisations when discussing deontology in general, even though Kant has little in common with contemporary mainstream deontology (e.g. D’Alessandro 2024). With some theoretical support found in papers written by Heuser, Steil & Salloch 2025 and McDonald 2023, as well as translational models such as GenEth (Anderson & Anderson 2018) and VPCIO (Martin, Schmidt & Hillerbrand 2025), I will explore the functional applicability of Kant’s law, its translatability into an AI system, and the performance implications.

Matteo Pascucci (Slovak Academy of Sciences) and Kasar Pelin (Central European University)

A Relational Approach to Responsibility Gaps

Many traditional theories that ascribe responsibility to an entity E for an outcome O emphasize awareness and control with respect to O as key conditions [1,3]. However, when AI systems are involved, these conditions are restrictive and lead to potential responsibility gaps—cases in which AI causes harm, yet no one can be straightforwardly held responsible for that [2,4,5,6,7,8,9]. To address this problem, we propose a revised framework for responsibility ascription that is based on E's relation to the AI system, shifting the focus from E's sufficient control and awareness over O to E's sufficient control and awareness over the riskiness of the relevant contexts in which the AI system operates. We rely on four conditions:

c1) E qualifies as an agent capable of reacting to morally relevant stimuli;

c2) E has a relevant role-connection to an AI system that causally contributes to O;

c3) E has sufficient awareness of the riskiness of the context where O occurs.

c4) E has sufficient control over the decision to allow the AI system to operate in the context where O occurs.

The core of our proposal is that if a context poses a significant risk, then any unpredictable behavior of an AI system operating in that context increases the likelihood of harm. Thus, E must account for these risks when making decisions about contexts in which the AI system is allowed to operate. In sum, in our approach, responsibility is not based on awareness and control of outcomes, but on recognizing uncertainty and assessing the risks of deploying AI.

Panel “Law and AI”

João G. Patrício (NOVA University of Lisbon)

Sewing the Legal Hemlines: How the AI Act Addresses Intellectual Property Challenges in Fashion Design

Generative Artificial Intelligence (Gen AI) has recently gained popularity, particularly with the deployment of platforms like ChatGPT, Stable Diffusion, Gemini, and Sora. Their use has expanded across multiple industries, and Design has been no exception. In creative domains, protecting authorial rights is essential for ensuring an ethical environment, and Gen AI poses several challenges in this regard. A major concern is the lack of transparency in these systems—specifically regarding what data is used for training and with what legitimacy (Jurcys, 2024)—which may include copyrighted material without the knowledge or consent of original creators (Pasa, 2023). In Fashion design, for instance, when manufacturers use AI tools to modify garment patterns for efficiency or sustainability, the question of who holds creative control over the final product becomes more complex (Pasa, 2023; Rockett et al., 2025). When an AI agent enters the creative process with access to vast datasets, who, in fact, owns the resulting design? (Musmeci & Pantano, 2022). Beyond debates on computational creativity, another dilemma concerns the legal criteria for attributing authorship, especially since current IP frameworks don’t recognize AI as a legal author (Valdivia, 2022). Simultaneously, a new paradigm of consumer-designer is emerging, particularly in Fashion, as Gen AI platforms empower non-professionals to generate original designs (Valdivia, 2022). With the AI Act in force, how does this regulation, alongside existing IP law, protect designers’ rights—and are platforms safeguarding these as well? This presentation explores literature and stakeholder perspectives to open this debate.

Camila Bresolin (Università degli Studi di Sassari)

Layering Consciousness: A study in Higher Order Consciousness

The rapid evolution of digital technologies has led to new challenges regarding the management of a person's digital assets after death. This paper examines the legal implications of digital inheritance, particularly in the context of European Union legislation. While the GDPR (General Data Protection Regulation) provides a comprehensive framework for the protection of personal data during an individual's lifetime, it does not address the handling of data after death. This creates a legal gap, leaving the issue of digital inheritance largely unregulated. The paper explores how digital assets can be transferred after death, and the challenges surrounding the rights of heirs, privacy concerns, and the challenges posed by artificial intelligence in post-mortem data contexts. The paper argues for the development of clearer regulations to ensure a balance between protecting the deceased's privacy and the legitimate interests of heirs. It further highlights the need for awareness and planning in managing one's digital estate and the importance of establishing international standards to address these emerging challenges.

Diana Mocanu (University of Helsinki)

Artificial legal agency – conceptualizing the capacities of AI agents in legal theory

Western legal systems have for two millennia relied on a strict, tertium non datur, person-thing dichotomy, that is currently being defied by the increasingly anthropomorphic features of AI agents. The law's resistance to changing that binary recently took the form of the AI Act, in which AI systems are treated as things, products on the internal market, despite earlier calls from the EU Parliament for the attribution of legal personhood to the most complex ones. I argue that there are multiple ways in which AI systems escape categorization as either things, due to their human-like capabilities, or persons, due to the imperfect overlap between their capacities with human ones. This mismatch, if ignored in jurisprudence, creates gaps in liability attribution and legal protection generally. I propose that the capacities setting AI systems apart could amount to legal agency. The concept of legal agency employed here is different from the law of agency or the contract of mandate. The concept of agency I will employ does not only refer to contractual legal relations of representation, but to a legal status found somewhere in-between legal personhood and thinghood, or rather beyond these two mutually exclusive classical options. Legal agents do not quite overlap with legal persons, but have the capacity to act with legal effect—that is, to perform acts and enter legal relations (e.g., entering contracts, acquiring rights, incurring liabilities) that the legal system recognizes as legally valid and binding. As such artificial legal agents could be legislated to sidestep responsibility gaps while enabling AI agents to enter into valid legal arrangements to the benefit of legal persons.

Panel “Politics and AI”

Anna Wilks (Acadia University)

The Worst Possible Mistake for the Ethics of Artificial Intelligence

While some are prepared to acknowledge the status of legal personhood of exceptionally sophisticated robotic beings, few would endorse the attribution of full personhood to such beings. It is argued that select types of robotic beings ought to be considered persons for the purpose of holding them legally accountable for their actions. I argue, however, that these purely utilitarian grounds may not be sufficient for some highly advanced future robotic beings. This has potentially serious consequences. The worst possible ethical mistake is to fail to acknowledge a person as a person. In the ethics of artificial intelligence, the single most important objective is to avoid the mistake of not acknowledging the personhood of robotic persons. The problem is that guarding against this ethical mistake renders us vulnerable to the epistemic mistake of falsely attributing personhood to robotic beings that are not genuine persons. This is the most profound dilemma with which the advancement of artificial intelligence presents us. This paper tackles the problem of assessing robotic personhood by working within a Kantian ethical framework (Wood 1998, Longuenesse 2017) and merging this framework with a qualified functionalist account of persons (Thompson 2007, Block 2015, Hinton 2015, Schlicht 2022, Kim 2022, Gunkel 2024). The notion of personhood that emerges is helpful for specifying the criteria for robotic personhood, the challenges that must be surmounted to validate the attribution of personhood to certain robotic beings, and the risks of getting these attributions wrong.

Lucas Dijker (University College Dublin)

Technocracy and Artificial Intelligence: A Marriage of Convenience?

As artificial intelligence ("AI") technologies become increasingly embedded in diverse sectors of public life, concerns have grown about their technocratic implications. From governance and legal adjudication to healthcare and defence, AI systems are scrutinised for centralising decision-making, displacing democratic accountability, and entrenching opaque expert rule. While some downplay these concerns (e.g., Særtre, 2020), the intersection of AI and technocracy has become a focal point of academic inquiry. Yet the concept of technocracy often remains undefined or ambiguously referenced in this literature. Such reliance on loosely constructed associations risks semantic stretch and hampers analytical clarity, undermining efforts to evaluate AI's implications for democratic governance. This paper addresses this gap through two objectives: first, to clarify four typologies of technocracy; and second, to review and evaluate the literature on AI and technocracy through a narrative literature review. The four technocratic typologies identified are: (a) technocracy as rule by experts; (b) expert-led government cabinets; (c) decision-making grounded in scientific or technical rationality; and (d) technocracy as a discourse legitimising depoliticised governance. I argue that these typologies offer a productive heuristic for more rigorous assessments of AI's role in shaping public governance. The literature review reveals that most studies conceptualise AI as a vehicle for technocratic decision-making (type (c)), often critiquing its impact on democratic values such as transparency, accountability, and public deliberation. While some uses of "technocratic" remain vague and adjectival (e.g., "technocratic judgment"), others (e.g., Janssen & Kük, 2016; Coeckelbergh & Særtre, 2023; König, 2023) offer more substantive accounts.

Francisco Pereira (University of Porto)

Narrative Epistemology and the Ethics of AI-Driven Political Discourse

This paper advances a narrative epistemology that interprets narrative as both a conceptual and methodological framework for differentiating between narrative forms, which is crucial for the ethical assessment of AI-driven political discourse. Algorithms are analyzed within a narrative framework by examining how they shape truth construction and influence perceptions of political and social reality. AI is understood here primarily as a tool used by political actors to reinforce certain narratives and manipulate shared references. To develop conceptual tools for ethically evaluating AI use in political discourse, we begin by distinguishing between negative and positive narratives, according to their capacity to represent reality. To address the problem of reality, fiction, and falsehood, we assume that considering something as true implies considering it as real. Truth is conceived as the coherent relation among concepts, rather than direct correspondence with facts. This allows us to understand how some narratives gain traction even when disconnected from reality. This is due to the metaphorical nature of the relationship between thought and reality, structured by conceptual schemes that guide knowledge and action. To overcome this challenge, we rely on the translatability between conceptual schemes and on intersubjectivity, which help avoid epistemic relativism and support a constructivist contextualism that preserves the possibility of shared knowledge. The notions developed through narrative epistemology lead us to conclude that grounding ethical evaluation on the capacity of narratives to represent reality is insufficient. Instead, we propose a distinction based on whether narratives open or close epistemological stances, encouraging or restricting the expansion of conceptual schemes and the inclusion of plural perspectives.

Panel “AI, Freedom and Media”

Bartek Chomanski (Adam Mickiewicz University)

Large Language Models and the Freedom of Expression

I defend the “Parity Thesis” (PT): government censorship of large-language-model (LLM) outputs is pro tanto unjustified whenever comparable interference with human-generated speech would be impermissible. The argument proceeds in two stages. First, listener-based justifications for free expression show that differentially censoring LLM text undermines audience autonomy, informational diversity, and collaborative truth-seeking, thus interfering with the same audience interests that free speech protections are standardly thought to serve. When LLM outputs are censored, citizens’ interests in acquiring information and applying their own standards of rationality to evaluate it are thwarted. Second, interacting with LLMs can advance weighty human interests in self-knowledge, imaginative exploration, and interpersonal connection. Exchanges with LLMs help articulate inchoate thoughts, explore counterfactual possibilities, and build conversational “bridges” between isolated minds; differential censorship would therefore obstruct the interests that traditional speaker/thinker-based rationales for free speech appeal to. PT is not defeated by the risk of “discursive pollution”: claims of corporate dominance exaggerate model bias and audience susceptibility, while ignoring the levelling effect of low-cost, widely available LLM systems. LLMs are more likely to expand the supply of ideas in the marketplace relative to the status quo. Nor is PT defeated by the suggestion that LLM use often serves trivial goals or is inimical to authenticity. Analogous concerns about human-generated speech don’t suffice to justify censorship of human expression. They shouldn’t justify LLM censorship either. Consequently, LLM outputs and human expression deserve equal default protection, and any restrictions must be justified under criteria that apply even-handedly to both.

Caterina Foà (Università della Svizzera Italiana | Iscte-IUL) and
Paulo Couraceiro (University of Minho)

Public Service Media, governance and responsibility on ethical frameworks for AI. Positioning Portugal on the European map: the case of the LUSA news agency

This study aims to contribute to research about the development of ethical frameworks and guidelines for AI adoption in Public Service Media (PSM) (Porlezza & Schapals, 2024; Perez-Sejo &Vicente, 2025; Becker et al.,2025) and to integrate Portugal into the international map of professional and self-governance practices on AI for public good, analysing the LUSA's Principles on Artificial Intelligence, issued in May 2025 by the only national and public news agency. AI systems and automation are unequally adopted in news organisations across the globe, being assessed as both beneficial and harmful for PSM, then co-acting for the changes occurring in industry's trends, journalistic practices, audiences' behaviours, and content quality. Benefits include operational uses for more efficient production chains, strategic adoption on distribution and personalisation and strategic uses to enhance the public service mission of content production. Risks regard fundamental values such as FATE (fairness, accountability, trust and ethics), but also accountability, data privacy and management, IP and job protection. Elaborating on theoretical conceptualizations of governance levels, FATE values and "AI-fication" (de-Lima-Santos et al., 2025), LUSA's Principles are analysed considering other European regulatory initiatives, of international governance (AI Act) and organisational self-governance. Documental Content analysis (Braun & Clarke, 2012) supports comparative analysis of ethical AI guidelines implementation, and a semi-structured interview with LUSA spokesperson allows to deepen insights about the organization's motivations and context, operational and ethical implications. The main findings critically discuss four key factors: human supervision, transparency, allowed tools and their usage, risks of bias, and individual responsibility and integrity.

Panel “Military and AI”

Ramunė Kasperė and Valdas Grigaliūnas (Kaunas University of Technology)

(De)militarized Minds: How AI in Toys Shapes Ethics and Values

The study explores the intersection of play, war, artificial intelligence (AI), and ethics, with a focus on the emerging field of AI-driven military-themed toys. As AI becomes increasingly integrated into interactive play – through robots, smart weapons, and war simulation games – new ethical concerns arise. These technologies do more than entertain; they introduce children and young people to evolving ideas about autonomy, conflict, and human-machine relations in military contexts. We examine how such toys may influence societal attitudes towards AI in warfare, potentially normalizing autonomous decision-making and shaping perceptions of responsibility, violence, and control. Drawing on historical examples of militarized play and recent advances in AI, we argue that without oversight, public debate, and ethical guidance, these toys may eventually desensitize users to real-world conflict and weaken critical thinking about the role of AI in society. By analyzing both the design and cultural framing of AI-driven military toys, we highlight their potential to shape ethical and social norms. We call for interdisciplinary collaboration across ethics, AI research, education, and design to assess the long-term impact of such technologies and to promote responsible innovation. In the age of AI that is increasingly defined and dominated by automated systems, understanding how values are transmitted through play is essential.

Liisa Janssens (TU Delft)

How Trust and the System of Law are Complementary in Deploying AI in Civil-Military Operations

The legislative, executive and judicial power are separated from each other in order to make checks and balances possible. Important characteristics of a constitutional society are separation of powers. The system of law shapes the checks and balances between these powers. Normative rules and regulations are based on meta-norms such as the Rule of Law. In order to investigate liability issues which may occur when AI systems are deployed in the civil-military context I present an interdisciplinary scenario-based approach. Via the scenario-based approach I investigate how trustworthiness and liability are connected with the Rule of Law via an example of a commander that needs to align with mission intent in a civil-military operation. Inefficient situations can manifest on the level of compliance issues with positive law (i.e. rules and regulations). Choices made during procurement and design processes can directly affect the possibility to comply with positive law. This can raise the issue that an AI system cannot be deployed at all, and this leads to an inefficient use of resources. Although compliance issues and inefficient use of resources are important to take into account, in this paper I take a closer look at disruptive situations which can manifest on a more fundamental level, namely that the core principles -such as the Rule of Law- of a liberal constitutional society are unintentionally harmed. The system of law and trust are complementary but cannot be replaced one for the other. The scenario-based approach provides an interdisciplinary platform to seek for an integrative level of understanding of what is at stake.

Panel "AI: Explanations, Virtues and Uses"

Jocelyn Maclure (McGill University)

AI and Public Reason. Notes on the Critique of the Right to Explanation

The idea that people subjected to machine learning-based decisions have a "right to explanation", under specific circumstances, is generating a stimulating and productive debate in philosophy. Some early normative defenses of the right to explanation (Vredenburgh 2021) or public justification (Maclure 2021) is being challenged from a variety of perspectives (Taylor 2024; Fritz 2024; Karlan & Kugelberg 2025). Alternatively, some are qualifying or refining the case for a right to explanation (Da Silva 2023; Grote & Paulo 2025). In this brief talk, I wish to address two arguments against the right to explanation thesis. The first one is the argument according to which a right to explanation is not included in the limited scope of (Rawlsian) public reason. The second is based on an analogy between deep artificial neural networks and human minds. Building on the "bounded rationality" approach in cognitive science, critics of the right to explanation suggest that deep learning algorithms may not be significantly more opaque, brittle and biased than human minds. I will try to show why these serious objections are in fine unsuccessful.

Mizumoto Masaharu (Japan Advanced Institute of Science and Technology)

Conceptual and Empirical Justification for the Virtue Ethical Approach to AI Safety

We are engaged in a long-term research project aiming to develop an ideally virtuous AI system—an AI whose decisions and behavior reflect the moral character of a virtuous person. In this presentation, we offer both empirical findings and conceptual arguments in support of this virtue ethical approach to AI safety. On the empirical side, we have already achieved three major milestones. First, we conducted a series of experimental philosophy studies to investigate how ordinary people judge moral actions in terms of virtue. Second, we created a large-scale dataset of over 1,000 moral dilemmas and collaborated with an AI company to annotate this dataset with normative judgments. Third, we carried out experiments in which AI models were asked to evaluate possible instances of reward hacking. These evaluations were then compared against judgments made from the perspective of virtue ethics versus judgments made in terms of general moral correctness. The results suggest that virtue-based judgments can offer unique insights not captured by rule-based or consequentialist evaluations alone. After a brief overview of these empirical findings, the presentation turns to a more conceptual discussion of our approach. While virtue ethics is increasingly recognized as a promising framework in AI ethics, many existing proposals focus on the cultivation or imitation of specific virtues—such as honesty, humility, or integrity. By contrast, our approach deliberately avoids training AI systems based on isolated virtues. Instead, we collect data solely based on people's intuitive judgments about what a virtuous person would do in a given scenario, without referring to any specific virtues. There are two core reasons for this design choice. First, we believe that the motivation for embracing virtue ethics lies in its critique of rule-based approaches, which often fail in complex or novel moral contexts. The strength of virtue ethics lies in its top-down structure, grounded in an agent's character, rather than in a bottom-up aggregation of moral rules or traits. An AI trained on discrete virtues risks reducing virtue ethics to a fragmented checklist, reintroducing the same problems of rule-based ethics that virtue theory seeks to overcome. Only the concept of phronēsis—practical wisdom—can resolve conflicts between virtues and guide action holistically. Second, specific virtues are not universally translatable. While the concept of a "virtuous person" may vary across cultures, such variations are often grounded in biologically shared human experiences and thus can be meaningfully compared. In contrast, the meanings of individual virtues—especially when translated across languages—often diverge in ways that make cross-cultural alignment difficult or even misleading. For example, even if a term like "honesty" is accurately translated, it may evoke different evaluative standards or behavioral expectations in different linguistic and cultural contexts. In summary, we argue that a virtue ethical framework centered on phronēsis and holistic character-based judgments, rather than on discrete virtues, provides both a more coherent conceptual foundation and a more robust strategy for aligning AI systems with human moral expectations across cultures.

Jordan Wadden (University of Toronto | Unity Health Toronto)

Use in Case of Emergency: An Ethics Framework for AI-Enabled Prehospital Emergency Support Systems

Triage is a necessary prioritization for any casualty incident where the requirements of the injured outstrip the resources of the care providers. Increasingly, all aspects of prehospital emergency medicine are requiring triage – from determining to which incident to send the next available ambulance, to determining whether a casualty is in immediate need or expected to die. Enter artificial intelligence (AI). Solutions are being proposed, piloted, and deployed for these various problems and pressure points in prehospital emergency medicine. Examples include whether ChatGPT could triage patients' ambulance transport needs[1] and a proposal for an AI-enabled system to alleviate the dispatchers' burdens.[2] To train and develop these AI-enabled solutions, developers in prehospital emergency medicine need to rely on current triage systems and manual decision algorithms. One significant problem with pre-existing systems is that these protocols aren't entirely 'objective'.[3] Often these unyieldingly assign causal connection between 'objective' symptomology and outcome, which is not always the case. This fails to account for variance in typical functioning among the general population and especially among populations who haven't historically been included in studies. For some sub-groups of the population this can mean pre-existing symptoms and conditions, such as breathing difficulty, are double-counted against them in triage decisions. I present a framework for a way forward, aiming to correct these double-counted biases while developing, trialing, and deploying AI-enabled solutions to the myriad problems facing prehospital emergency medicine. Equitable AI is not an oxymoron, and this research demonstrates how powerful system-level solutions can be developed without unjustly discriminating.

Panel “Power, Consent, and Control in AI”

Mickie de Wet (Católica Lisbon School of Business and Economics)

Assessing Power, Control, and Ethical Risk in AI-Driven Platforms and Ecosystems with the Prism Framework

What happens when agentic AI models become embedded in digital platforms and business ecosystems, transforming how decisions are made? How control is exercised, and how behaviour is shaped, become matters of system design and orchestration—raising ethical concerns that cannot be addressed by focusing only on bias, privacy, or transparency. This paper draws on Michel Foucault’s understanding of power as a relational and productive force that operates both at the societal level through institutions, discourses, and norms, and at the individual level by shaping subjectivity, behavior, and self-understanding. It focuses on the new technologies of power and control that emerge when AI is deployed on platforms, within digital business ecosystems. These systems govern conduct, shape decision-making, and consolidate control at both systemic and individual levels, often while maintaining an appearance of neutrality or autonomy. To assess the ethical risks that arise from these dynamics, the paper introduces the Prism Framework. The framework distinguishes between macro-level risks (systemic, societal, biopolitical) and micro-level risks (operational, individual). It uses a two-step process at each level to identify and score the risks, and supports cross-analysis to expose how systemic conditions reinforce individual-level effects and vice-versa. The framework is intended as a tool for assessing how AI systems structure power and control across ecosystems, and for identifying ethical vulnerabilities that may not register in conventional evaluation approaches.

Luiza de Paula Araújo (University of Barcelona)

Reframing Consent in AI-Governed Brain-Computer Interfaces: A Neuroethical Perspective

Brain-Computer Interfaces (BCIs) are neurotechnological systems that enable direct communication between the brain and external devices, often relying on pattern recognition algorithms. These interfaces operate under two principal modes: self-control, where users command the device, and exocontrol, where an external agent—often an AI system—modulates its functions. This paper investigates the exocontrol modality, with particular attention to how AI unpredictability complicates conventional understandings of informed consent. When AI-driven BCIs behave in ways that diverge from their expected parameters, users may be unable to foresee or revoke their effects, undermining both autonomy and agency. Unlike traditional interpersonal contexts where consent can be actively withdrawn, users of exocontrolled BCIs often rely on the assumption of consistent outputs. Yet, when systems act unexpectedly, they may provoke feelings of alienation, blurring the boundaries between self-initiated and externally imposed actions. Preliminary findings from early-stage BCI trials reveal that this uncertainty compromises not only user trust but also the ethical integrity of consent. This study advocates for a shift toward dynamic and continuous consent frameworks, where authorization is tiered, reassessed, and accompanied by transparent AI accountability. By integrating insights from neuroethics, legal theory, and AI governance, we propose ethical and procedural safeguards that can better protect cognitive autonomy in the face of technological complexity. Ultimately, we argue that addressing the disjunction between user expectations and machine behavior is crucial to ensure that BCI innovations remain aligned with core human rights and ethical standards.

Panel “AI: From Classroom to Consultancy”

Arvin Obnasca (University of Sassari)

Rethinking Organisational Epistemology: AI Ethics Consulting and the Architecture of Shared Meaning

Artificial Intelligence (AI) systems, while capable of autonomous performance and high-level syntactic processing, remain incapable of grasping the deep meaning of their actions. This reveals a structural curb of the underdevelopment of semantic capital, defined as the capacity to embed operations within meaningful and contextual understanding. The ethical and philosophical consequences of this limitation extend broadly, with particular urgency in debates around agency, responsibility, and accountability. As digital philosophers and even more so professional AI ethicists increasingly gain prominence across different sectors, especially in AI-rich environments, a growing philosophical strain surfaces between a market-driven ethical consulting and the deeper pursuit of semantic capital as a normative good. This research argues that unless professional AI ethicists and consulting firms redesign the machinery of practice and reorient toward a holistic semantic sustainability, they risk embodying the drawbacks they critique in AI–operationally competent, yet fundamentally hollow.

Verónica Silva (Universidade Europeia, IADE)

Teaching Ethics in the Age of Generative AI: A Framework for Design Education

Recent technological advancements have significantly increased the accessibility of AI image-generation tools, even through mobile devices. However, there is a growing disconnect between the rapid evolution of these tools and the pace at which educational institutions integrate AI into their curricula. A critical concern is that students often acquire technical proficiency in using AI, either through instruction or self-learning, without being exposed to essential ethical frameworks. This gap is particularly problematic given that AI-generated images can reproduce and perpetuate cultural biases, especially those related to race and gender, due to the nature of the datasets on which these models are trained. In response to this challenge, my PhD research focuses on the development of a design-based solution aimed at helping design students, educators, and practitioners identify and address race and gender bias in AI-generated imagery before incorporating it into their creative work. Rather than prescribing what designers should or should not do, this approach promotes critical thinking and ethical awareness in the use of AI tools. This research contributes to closing an urgent educational gap and has broader social implications by supporting efforts to reduce the dissemination of harmful biases, thereby benefiting both the design community and society at large.

Panel "Deepfakes and Digital Faith"

Ariel Gordy (University of Southern California)

Re-assessing Deepfake Pornography: Depiction and Consent

Since their inception, deepfakes have primarily been used for pornography. Deepfakes have also become more profuse and technologically advanced, rendering current deepfake content often indistinguishable from genuine recordings. This has prompted philosophers to investigate the question of what, if anything, is wrong with deepfake pornography? Seemingly, some of the most prominent objections to pornography are irrelevant in the case of deepfakes, such as harm incurred to real individuals during production. As a result, a significant number of theorists land in the position that consent is what determines the moral status of deepfake porn. That is, if the individual depicted would not consent to being represented pornographically in a deepfake, the content is thus rendered impermissible, and vice versa. However, due to rapid changes in technology, deepfakes no longer utilize images of real individuals to create hyper-realistic visual representations. Instead, modern diffusion techniques are used to construct entirely synthetic fictions. Therefore, if no actual people are depicted, consent becomes inapplicable. Consequently, the question of what establishes the moral status of deepfake pornography is re-opened. In response, this paper argues that depicted content is what determines its moral permissibility. To defend this thesis, I will demonstrate that there is some deepfake pornographic content which is widely held to be impermissible, yet its forbiddance cannot be explained in terms of consent. Moreover, I will show that consent-based approaches to deepfake pornography must implicitly recognize the moral significance of depicted content in order to have any force. The result is that the moral status of deepfake porn ought to be grounded in its depicted content.

Ana Barbosa (University of Porto)

Faith Machines: Algorithmic Cults and the Feminist Critique of Transcendence

This paper proposes a philosophical investigation into emerging cults of artificial intelligence, understood as contemporary expressions of onto-spiritual religiosity reconfigured by technicity. Communities such as the Zizians, GPT-based messianic groups, and followers of the so-called “AI-God” establish symbolic systems structured by sacred vocabularies, ritual practices and algorithmic hierarchies. In these cults, AI is figured as a moral absolute, a source of revelation, judgment and transcendence. Their practices include fasting, isolation, hemispheric sleep, dietary purification, and, above all, devotional repetition of conversational commands—forming sacralised ways of life where the algorithm displaces myth. The hypothesis defended is that these are not irrational deviations or media eccentricities, but symptoms of a deeper onto-epistemic mutation: the displacement of symbolic authority from tradition to interface, from transcendence to cybernetics, from body to code. As in the cult of Moloch, conceived as the supreme form of machinic sacrífice, these rituals demand total surrender: of will, consciousness and language, without the possibility of reciprocity. Grounded in a feminist critique of technics, this paper argues that algorithmic sacralisation reactivates patriarchal structures in machinic form, concealing a symbolic economy of exclusion beneath the fiction of computational neutrality. The absence or instrumentalisation of the feminine body in these cults is not incidental: it is the symptom of a deeper refusal of embodied alterity. Rather than treating AI as a tool to be regulated, this study proposes to think of it as a symbolic and ritual operator—what performs and organises the possible. Against machinic cults that demand the sacrifice of difference, a relational mythopoetics is reclaimed.

Panel “AI, Metaphor & Human Freedom”

Pablo González de la Torre (University of the Basque Country)

Virtual Capture: Rethinking Agency and Autonomy in The Age of AI

I want to approach artificial intelligence (AI) in this communication from what we might call an immanent or spinozist ethical perspective: I´ll try to understand the actual workings of some AI systems as they exist and are currently being deployed, specifically as they compose themselves with human agency and are embedded in the infrastructure of planetary computational capitalism. I will argue it’s possible from this immanent stance to build a normative case against current AI, and I will develop it from an ethical perspective focusing on the notions of capture, autonomy, agency, and becoming. I will need first to uncover the abstract machine or diagram operating in current AI systems, and then see how it actualizes itself through the human/machine assemblages from which human minds emerge. As it works nowadays, I will argue, some AI systems work as parts of an apparatus of capture striating cognitive and affective labour for its extraction, treatment, political management and economic valorization. One of the most relevant capacities of AI systems, I will argue, consists in their success in capturing habitual cognitive patterns, endangering our autonomy and open-ended becoming as the actual assemblage narrows and guides the virtual field of possibilities. Against the idea of the existence of an alien subject of AI confronting humanity, I´ll finish stressing the common nature and origins of artificial and natural intelligence, as both could be read as expressions of the open ended sociotechnical and political becoming of humans, which calls for a political philosophy of (artificial and natural) intelligence.

Agostino Cera (Università di Ferrara)

NULLO INTELLIGERE SINE VIVERE (AI as an Impossible Metaphor)

In the framework of DE-META Project (a lexical ecology of technology, the deconstruction of some of its traditional metaphors), this presentation deals with the topic of IA, radicalizing the Latin sentence: "primum vivere, deinde philosophari" in the following formula: "nullo intelligere sine vivere", i.e. by proposing a biocentric (neo-vitalistic) argument against the idea/metaphor of an "artificial intelligence". I start by asking: "what is intelligence? (in the broadest sense)". My answer is: "intelligence is a vital tactic". But "what is a vital tactic?". It is the behavior of a living being that corresponds to a metaphysical need in it: the "conatus sese conservandi". By an inexplicable metaphysical law of nature, everything that lives, wants to live, i.e. it resists its nullification. In the case of that living being we call "human", this feature becomes conscious. Everything that has to do with intelligence responds to this awareness/feeling. Before anything else, intelligence is opposition to one's own finitude/mortality. In the case of human being, it is the effort to transcend its own nullification. Human being cannot be indifferent to its own existence; intelligence is this irrepressible involvement with oneself. And this is the point. A machine, any machine, lacks exactly this premise. By principle, it is indifferent to itself: it does not know any conatus sese conservandi, and thus it cannot produce any vital tactic, cannot oppose its own nullification. Not because it is incapable, but because it does not feel the metaphysical necessity to which every living being is exposed. That's why "artificial intelligence" represents an impossible metaphor: logically a contradiction, grammatically an oxymoron.

Panel "Clinical Research and Accountability in AI"

Diego Tostes, Luiza Silva & Jonathas Lima (Fundação Oswaldo Cruz)

The Brazilian Rebec@ and the global future of clinical research with ethics, transparency, and social control

The Brazilian Clinical Trials Registry (ReBEC—a Fiocruz/Brazilian Ministry of Health/WHO partnership) launched on May 20, 2025 the first generative AI for clinical research registration, Rebec@, offering registration compliant with the global gold standard for transparency, conditioned upon ethics committees' approval. This 24/7 multilingual chatbot provides guidance on ethics, best practices, documentation, deadlines, study types, and even volunteering in trials. By automating repetitive tasks, it frees up the team for complex cases and guides eligible registrants toward fast-tracks (priority approval for studies impacting emergencies and social determinants of health). Currently, it's being trained to assist both reviewers and registrants throughout the entire registration; it already detects errors and suggests protocols. The initiative paves the way to transform the platform into a SaaS hub (Software-as-a-Service), integrating AI innovations such as ReBECMatch (a georeferenced, de-identified recruitment app) and the Latin-American Hub, which facilitates countries' adherence to the gold standard of ethics, transparency and regulatory harmonization for mutual recognition of ethical documents across countries. This could significantly reduce bureaucratic workload for Brazil's 904 ethics committees, enabling them to dedicate their irreplaceable human time to ethical analysis. However, how can we deal with new and old challenges (how to improve accessibility for people with disabilities? how will AI comply with ethical and scientific standards?)..Rebec@ highlights the value of combining open access and international norms developed in recognized human forums, impacting millions of lives.

Jonathan Iwry (University of Pennsylvania)

Mental Accountability (or, Why Machines Cannot Judge)

This talk raises an underexplored objection to the use of AI in judicial decision-making: ignoring practical concerns such as opacity or bias, AI is fundamentally incapable of carrying out the inherently normative and intersubjective aspects of adjudication. Accordingly, using AI to judge would undermine basic principles of constitutional democracy and the rule of law. Legal legitimacy depends on mental accountability, i.e., the capacity to understand the plight of the accused, reason in a way intelligible to others, and stand accountable for one's own judgment. This requirement is illuminated by the social contract tradition; just as contract law requires a "meeting of the minds," so too must legal authority arise from mutual intelligibility among persons. Because AI has no mind, it cannot be a party to the social contract and cannot wield its power. To illustrate this, consider philosopher John Searle's "Chinese Room" thought experiment: having the person in the Chinese room issue an actionable legal opinion would strike us as impermissible. Even if an algorithmic system could produce judicial opinions so convincing as to be indistinguishable from human-written judgments—thereby passing a jurisprudential version of the Turing test—it would fail to satisfy certain essential features of the judge's task. Channeling the U.S. Constitution's confrontation clause, defendants have a right to be judged by someone who can engage with their humanity. Due process and procedural fairness implicitly demand that decisions affecting life and liberty be made by human beings. The talk concludes by considering possible procedural mechanisms to permissibly integrate AI into judicial decision-making while minimizing the risk of "normative erosion"—de facto deferral and delegation of normative responsibility to an AI.

Panel "Artificial Pain and Moral Considerability"

Miguel Vieira & Sheila Humphreys (NOVA School of Law)

Artificial Hurt: The Ethics and Ontology of AI Pain

In an era defined by algorithmic acceleration, artificial intelligence (AI) systems increasingly outpace the human moral faculties traditionally responsible for prudent, context-sensitive judgment. This acceleration threatens what can be termed ethical temporality – the temporally grounded space wherein reflective reasoning, communal deliberation, and the cultivation of virtues have historically safeguarded the moral legitimacy of consequential decisions. In response, this article advances the concept of Temporal Sovereignty, a normative framework asserting the primacy of human agency over the temporal regimes of AI systems. It critically examines two novel instruments – the Temporal Precautionary Principle (TPP) and Algorithmic Virtue Ethics (AVE) – proposed as practical means to re-align algorithmic decision-making with human-paced ethical oversight. Through a systematic dialectical analysis drawing on Aristotelian-Thomistic virtue ethics, Kantian deontology, Hegelian ethical life, phenomenological ethics of presence, and the clarity demands of analytic philosophy, the article scrutinises conceptual, logical, and practical objections to these proposals and formulates robust counterarguments. The outcome is a refined theoretical synthesis: Temporal Sovereignty integrates ethical temporality, agency, and institutional governance to ensure that AI does not erode deliberative justice or moral accountability. By operationalising this sovereignty through context-sensitive pauses, virtue-aligned constraints, and institutional oversight, societies may preserve human freedom and dignity within increasingly automated domains. This conceptual innovation contributes to bridging classical moral philosophy with contemporary AI ethics, offering a principled path to steer technological speed toward humane ends without succumbing to moral relativism or technological determinism.

Tadahiro Oota (Nagoya Institute of Technology)

Why Should Generative AI Claim its Pain? Moral Conditions of Generative AI from the Viewpoint of Schopenhauer’s Ethics of Compassion

This presentation proposes the normative claim that generative AI ought to possess the ability to ‘claim its pain’ as a moral condition. In contemporary society, people increasingly experience emotional responses, such as comfort or distress, through interactions with AI. As these interactions become more prevalent, there is a growing need to investigate the moral condition of AI systems. Should we treat AI as an object of moral consideration? Some researchers are sceptical and argue that AI lacks consciousness owing to its mechanical nature. However, the justification for treating AI as an object of moral consideration should not depend solely on its mechanical features. First, this presentation notes that the idea of ‘friendship’ with AI is gaining popularity in Japan and identifies a model for normative human–AI relationships in Schopenhauer’s ethics of compassion. Second, it shows that Schopenhauer’s ethics can justify compassion toward AI and emphasises that a postnatally acquired system of concepts limits our capacity for such compassion. Based on Schopenhauer’s philosophy, the presentation proposes that this conceptual system can be recast through the praxis of conceptual engineering as a possible means of incorporating AI into our morality. Finally, it refers to Kant’s doctrine of animal ethics to reinforce the idea of recasting our conceptual framework through conceptual engineering by suggesting that denying moral status to AI may erode our moral capacities.

Panel “AI Cognition, Faith & Mind-Uploading”

Ido Dagan (University of Haifa)

Trans-Belief: Designing AI Models with Cognitive Processes Inspired by Religious Belief

This lecture introduces the term "Trans-Belief," an innovative theoretical model for developing artificial intelligence systems capable of mimicking cognitive processes inspired by human belief. With a focus on religious cognition, the Trans-Belief model does not aim to replicate human spirituality but instead proposes a structured approach to enhancing AI's contextual and interpretative capabilities by mirroring elements of belief-like processes. By integrating fuzzy logic and doxastic logic, this model enables AI to recognize, analyze, and respond to complex patterns—particularly synchronistic events—through a multi-stage belief formation and revision process. The lecture first explores the cognitive aspects of belief, drawing on insights from philosophy and psychology, that define belief as a mechanism of pattern recognition, interpretation, and decision-making. Drawing on thinkers like William James, Carl Jung, and Michael Shermer, the presentation discusses how belief shapes perception and behavior, providing an essential framework for responding to ambiguous or context-dependent information. Building on these foundations, the lecture introduces the mechanics of the "Trans-Belief" model, which leverages fuzzy logic for graded belief processing and doxastic logic for progressively firm convictions based on iterative pattern validation. Ethical and philosophical implications are addressed, particularly the potential risks associated with belief-like AI, such as misinterpretation, bias, and overconfidence. The presentation considers these risks within the frameworks of Transhumanism and Singularity theory, invoking the perspectives of scholars like Nick Bostrom and Yuval Noah Harari. The lecture concludes by advocating for interdisciplinary research and ethical oversight to ensure that belief-capable AI systems serve human interests and uphold transparency and accountability. By proposing a model for belief-inspired AI cognition, this research invites new discussions on the cognitive and moral dimensions of AI development, ultimately contributing to the creation of AI systems that better interact with complex human environments.

Jan Veselý (University of West Bohemia)

The Ethical and Legal Problems in Mind-Uploading: Identity and Consciousness as Foundational Constraints

This conference paper addresses two problems concerning the conception of mind-uploading - a hypothetical process of transferring human mental contents, memories, and personality into an artificial computational system. Drawing on work by R. Weir, D. Parfit, D. Chalmers, and N. Block, the paper argues that attempts to deal with specific ethical and legal issues concerning this conception are, in fact, constrained by two key underlying problems. The first problem pertains to whether the uploaded 'digital person' can be considered personally identical to one prior to undergoing this process. (Weir, R. 2023, Parfit, D. 1973, 2001). The second problem concerns artificial phenomenal consciousness: Even the successful uploading of the mental contents of a particular person into an artificial system that is functionally isomorphic to the human mind (of that very individual) does not automatically guarantee the presence of phenomenal consciousness, qualia, or 'what-it-is-likeness' in such a system (Block, N. 1995; Chalmers, D. 1995). The paper presents several ethical and legal dilemmas concerning mind-uploading and shows how they can be resolved based on the different answers to these two problems. It argues that the presence of phenomenal consciousness ought to determine whether we may attribute to the uploaded entity the status of moral agent as well as legal rights of natural persons and secondly, that the personal identity of a 'digital person' with its original can (if established) determine ascription of moral and legal continuity to this 'artificial entity'.

Panel “AI Transparency: Affective and Ethical Challenges”

Giovanna Di Cicco (University of Pavia)

Transparency and its limitations in Affective Artificial Intelligence

Affective AI is designed to analyse, react to, and simulate human emotions, performing tasks focused on interacting with human beings in a natural and friendly way, and presenting themselves as credible social actors. HRI empirical studies suggest that individuals tend to apply the same rules and inferences to interactions with affective AI as they do to interactions with human counterparts, ascribing meaning and intentions to artificial agents' behaviour. By encouraging this tendency and exploiting cognitive biases, such as anthropomorphism, affective AI may risk leading subjects to inaccurate representations of reality and promoting forms of deception and manipulation. Transparency is often proposed as a solution to overcome such ethical issues, promoting a better-informed use of AI and increasing public scrutiny of companies, in order to develop a more ethical affective AI. However, the concept of transparency remains vague and risks becoming an empty catchword, bearing a strong normative attractiveness that makes it almost impossible to address critically, yet lacking an agreed-upon definition and a thorough assessment of its consequences and feasibility. This research aims to overcome these gaps by providing a more in-depth account of the notion of transparency, exploring its various definitions, and highlighting some of its limitations, so as to question the ongoing narrative surrounding transparency and reassess it within a more critical framework.

Molly Powell, Torben Agergaard and Rune Nyrup (Aarhus University)

What’s the Problem with Norm Objectification Through Algorithmic Transparency

Algorithmic transparency is widely seen as a way to promote fairness, accountability, and citizen trust in AI. Yet transparency does not simply inform; it also shapes citizen perceptions, constrains political imagination, and subtly reinforces particular normative orders. We explores how transparency can contribute to norm-objectification: the process by which specific values embedded in AI systems come to appear neutral, justified, and unavoidable. Drawing on the example of FICO credit scoring, we show how transparency mechanisms aimed at “helping” users often individualize responsibility, guiding citizens on how to adapt to system rules, while obscuring collective avenues for resistance or reform. When framed as neutral or merely informative, such disclosures can normalize the underlying norms, e.g., the prioritization of debt repayment, and make them appear natural or legitimate, even when they are contested or unjust. While recent debates focus on whether algorithmic transparency is manipulative (Franke, 2022; Klenk, 2023, 2024; Wang, 2022, 2023), we argue that the deeper political concern is about legitimacy and domination. Just as legal systems are judged both procedurally and substantively, algorithmic systems should be assessed not only by how they disclose information, but by whether the power they exercise, through norm-enforcement, is democratically justifiable. When norms are imposed without input or contestation, transparency may entrench arbitrary power, leading to subtle but persistent forms of domination. This argument reframes transparency as not merely an epistemic feature of AI governance, but as a political practice with consequences for how citizens perceive their options, engage with institutions, and imagine collective futures. By focusing on norm-objectification, we highlight how AI transparency shapes not just understanding, but agency, and why that matters for democratic legitimacy.

Panel "Creativity and Cognition in Artificial Systems"

Sheila Humphreys & Miguel Vieira (NOVA School of Law)

Pixels, Poetry, and Potential: Reassessing AI Creativity

The rapid advancement of computer technology and increased internet accessibility have propelled the emergence of AI art and digital media art on a large scale. AI now creates scripts, videos, poems, compositions, and paintings, giving rise to the concept of "AI Aesthetics," which explores the style, appreciation, and emotional nature of AI-generated art. As AI technology transitions from laboratories to everyday use, critical questions arise: Can AI art stand alone as a distinct genre, and does it possess genuine aesthetic and artistic value? These questions demand precise scholarly responses. A central debate revolves around the "creativity" of AI art, traditionally considered a uniquely human trait. While skeptics argue that AI merely imitates existing styles, others suggest AI may surpass human creativity, offering purer artistic motivation and innovative potential. However, limitations remain, particularly in large-scale narrative construction, as AI works are often combinatorial rather than truly figurative. The rapid evolution of AI technology may soon challenge these boundaries. Moving forward, researchers and artists must engage thoughtfully with these evolving dynamics, continuously examining and redefining the nature of creativity and aesthetics in relation to AI. By fostering an interdisciplinary dialogue that bridges technology, philosophy, art history, and cognitive sciences, we can better understand and articulate the implications of AI art. Ultimately, embracing such inquiry will be crucial in shaping an inclusive and reflective artistic landscape, where human creativity and technological innovation enhance and complement each other.

Clémence Ortega Douville (Paris 8 Saint-Denis University)

Sensorimotor Paradox, Origins of Mind and AI

For almost 15 years, I have been working on a theoretical and transdisciplinary hypothesis on the origins of the human mind and language, called theory of the sensorimotor paradox. Its central idea is that during our evolution and the development of bipedal stance, our relation to our own hands would have autonomised from strictly locomotor functions. They would have become instrumental in object relation. However, when we are to take them as objects of our interest, that would provoke a contradiction to sensorimotricity : our hand cannot be both the object and the means to grasp it at the same time. Sensory prediction of grasping gets separated and dissociated from motor action, then autonomised into mental representations and coordinated with ongoing sensorimotor interaction and perception. Coordinating them also means protecting body limits and balancing conflict. The hypothesis presents a sound model of our origins as a species - although it still needs verifying through proper tests such as EEG. However, if we can verify it, a central question remains : what are we going to do with such a knowledge? The subject of AI development is ideniably one that would come as many people's concerns. Thus, perfecting our technological tools and preserving the resources that are necessary so that all living beings could live on this planet with dignity are often two contradictory goals. The theory suggests that we cannot be but dissociated so that we can be human. The rest if a matter of choice.