Joint IACAP/AISB Conference on Philosophy of Computing and AI

IACAP/AISB-25

Book of Abstracts

UNIVERSITY OF TWENTE.

TU Delft Delft University of Technology

IACAP International Association of Computing and Philosophy

AISB

Table of Contents

Abstracts within each category except the symposia are alphabetically sorted by the last name of the first author. For the symposia, sorting follows the sequence provided by the respective organisers. All Table of Contents entries are clickable links.

PLENARY TALKS

Relational Intelligence and Human–AI Relationships: Foundations, Technologies, and Ethical Dimensions

Philip Brey

In this talk, I will make the case that a new area is emerging within AI research and development: relational intelligence (RI). RI is concerned with equipping AI systems with the capacity to build, sustain, and navigate social relationships with humans. Defined as a transdisciplinary field, RI draws on technical domains such as affective computing, socially intelligent AI, social robotics, human-computer and human-robot interaction, and natural language technologies, while also incorporating insights from psychology, cognitive science, ethics, and sociology. The concept of RI extends beyond general social interaction to encompass specifically relational capabilities, including empathy, social interaction memory, moral cognition, reciprocity, trust-building, and relational maintenance and repair.

Currently, RI mostly exists as an applied field focused on product development. Products include social robots for companionship like Elliq and Moxie, mental health chatbots like Woebot, AI chatbot companions like Woebot, and virtual home assistants like Gatebox. In the future, RI is likely to include products that take on a wide range of social roles in a large variety of contexts. It is also likely to develop into an enabling field in A, and possibly even a foundational field.

In the talk, I make the case for RI as an emerging field in AI, and I propose conceptual foundations, including a statement of its core concepts, central aims, the capabilities that it models, and the relation of RI to other fields in AI. Next, I will discuss RI technologies and RI use cases, covering core product types, relational roles and functions, and key domains of application. I will do a forward-looking assessment of RI's future development. I will conclude with a discussion of social and ethical dimensions of RI and the need for robust ethical safeguards and governance frameworks to ensure RI's responsible development and use.

Using AI as Philosophical Method

Vincent C. Müller

We propose a new tool for the philosophical toolbox: "AI Philosophy". AI Philosophy systematically uses the example of AI with the aim to gain insights into genuinely philosophical problems. In this use, philosophy is not merely "applied" to AI, but the application generates insight in philosophy itself. Rather than asking whether an AI system has a particular property, we can use the confrontation with AI to ask what that property is, i.e. we loop through the formulation, application and evaluation of the theory. This method avoids anthropocentrism and gives us a new way of testing our philosophical theories. Given the wide range of features we can consider for AI systems, this method allows us to cover a wide range of philosophical issues, especially in the philosophy of mind, language, epistemology, and ethics. We outline the method, discuss some examples of its use, and consider objections.

Machine Learning in Science: Dimensions of Understanding

Emily Sullivan

More and more sciences are turning to machine learning (ML) technologies to solve long-standing problems or make new discoveries—ranging from medical science to fundamental physics. At the same time, the exact same modelling technologies are used across society ranging from determining what news we see on social media to fraud detection and criminal risk assessment. The ever-growing fingerprint ML modeling has on the production of scientific and social knowledge comes with opportunity and also pressing challenges. In this talk, I discuss how philosophy of science and epistemology can help us understand the potential and limits of ML used for science and society. Specifically, I will draw the themes regarding the nature of scientific modeling, understanding, explanation, and idealization.

INDIVIDUAL TALKS

When Predictions Are More Than Predictions: Self-fulfilling Performativity and the Road Towards Morally Responsible Predictive Systems

Markus Ahlers, Philippe van Basshuysen

In recent years, much attention has been paid to bias, discrimination and fairness in machine learning (ML) systems. Another ethical problem occurs, when performative predictions influence the outcomes they predict. Performativity has attracted attention primarily in the form of self-defeating prophecies, where the predictions undermine their own accuracy. The ethical dimension of self-defeating prophecies is widely recognized. In contrast, performativity in the form of self-fulfilling prophecies has remained largely unexplored, especially in computer science. But this also raises serious moral questions that need to be addressed.

For instance, risk assessment tools like COMPAS may predict high recidivism risk for a defendant, leading to their incarceration. This, in turn, may increase their recidivism risk due to disrupted lives and exposure to criminal environments, ultimately making the prediction come out true. While the prediction is accurate, it is so for problematic, self-fulfilling reasons, raising significant moral concerns.

In our talk, we highlight the ethical and legal dimension of self-fulfilling prophecies in relation to machine learning and call on the ML and ethics community to engage this form of performativity as well.

Could AI Assuage Loneliness? and If So, Which Kind?

Ramón Alvarado

Symons and Sanwoolu (forthcoming) suggest that given that an AI product could be available to many people simultaneously and without conventional social or physical restrictions, it will be unable to meet certain conditions – such as scarcity, uncertainty, and friction – that ground meaningful social connections. If this is true, then AI will be unable to have any bearing on or assuage loneliness.

In this paper, I argue that there is no such thing as 'addressing loneliness' simpliciter. There are distinct kinds of loneliness, and they are responsive to distinct kinds of interventions (Creasy, 2023; Alvarado, 2024). Hence, perhaps it proves more fruitful to ask which kind of loneliness could AI address, if any. I conclude by suggesting that as an epistemic technology, AI may very well be able to address epistemic loneliness (Alvarado, 2024) – a kind of loneliness that arises in virtue of the absence of epistemic peers with which to construct, accrue or share knowledge. This may be the case, however, only if we can deem AI as an epistemic partner (ibid) – a willing, able, actual, and engaging epistemic peer.

It Is Not a Camera! Radio Sensing Holography as a Disruptive Technology

Ciano Aydin, Stefano Savazzi, Sage Cammers-Goodwin, Sanaz Kianoush, Luca Posatti

The HOLDEN project investigates the ethical, social, and cultural implications of radio sensing holography, a groundbreaking technology that uses RF wave sensing and AI to create dynamic 3D visualizations of environments and track individuals. Unlike traditional cameras, holography enables "ubiquitous vision," seeing through walls and darkness. This capability raises critical concerns about surveillance, privacy, and autonomy while reshaping human-environment interactions.

Holography extends sensory capacities, reshapes perceptions, and actively influences behavior through AI-driven nudging and persuasion. Its technological gaze, unlike static observation, evolves into dynamic interaction, intensifying the feeling of constant visibility and subtly influencing decision-making. Applications include optimizing public crowd flow, delivering health prompts, and linking lifestyle factors to chronic disease risks via unobtrusive RF signal collection.

The research explores the integration of holography into a.o. smart homes, healthcare, and public safety, highlighting both opportunities for social connectivity and risks of inequality and privacy erosion. Case studies emphasise its dual potential to enhance convenience or reinforce power imbalances, depending on governance and design.

By advocating for privacy, accountability, and inclusivity, the project underscores the importance of responsible innovation to ensure holography fosters equity, autonomy, and human flourishing while transforming societal norms and human interactions.

Robots and Resentment

Don Berkich

The prevailing philosophical view seems to assert that fully ethical artificial agents require (at least) a trio of cognitive capacities – awareness, understanding, and free will – each of which presently exceeds either our engineering grasp or our computational modelling by no small measure. Consequently we shelve puzzles over, say, robots as fully moral agents in favor of the problem of designing moral normative constraints on their ranges of behaviour. Extending recent scholarship on the arguments PF Strawson offered in his influential "Freedom and Resentment", in this talk I argue that the nature of such participant reactive attitudes as resentment and gratitude are such that they apply to us regardless of whether we have awareness, understanding, or free will, and will apply to artificial agents regardless

of whether they have the same or similar cognitive capacities as human persons. Either this is a reductio on Strawson's argument, or fully ethical artificial agents are much nearer on the horizon, from an engineering standpoint, than current philosophical consensus allows.

Epistemic Type Mismatch

Yves Bouchard

When considering a fragment of what an epistemic agent knows, one can collect this information in a set in which all epistemic items can be represented as assertions. Such assertions, in turn, can be conceived as facts or rules for asserting facts. A set that contains both facts and rules is a knowledge base (Levesque and Lakemeyer 2000; Gelfond and Kahl 2014). A knowledge base can be queried to verify whether an assertion about a fact is true or not, and such a process can be assimilated to an inference. Now, if the knowledge base is populated with assertions of different types, i.e. epistemic items of different types such as perceptual knowledge and a priori knowledge, then under what conditions exactly can one perform an inference involving different knowledge types? Is the inferential operation exposed to the possibility of an epistemic type mismatch?

In the first part of the talk, I will analyse the inferential epistemic-type-mismatch problem and its consequences on a theory of knowledge. In the second part, I will present a way to cope with this difficulty in epistemic logic and the requirements for an adequate theory of knowledge.

From Simulating Towards Duplicating the Brain - the Case of Neuromorphic AI

Johannes Brinz

In 1980, John Searle argued against the ubiquitous AI optimism of his era that in order for a system to truly understand, it needs to duplicate rather than simulate the causal powers of the human brain. Despite the common recognition of the simulation-duplication distinction in philosophy, its articulation remains sparse, as neither Searle nor later commentators explicitly delineate what constitutes brain duplication versus mere simulation. This question became ever more pressing with the advent of so called ”neuromorphic computers”, i.e. hardware – chips designed to mimic neural structures, comprising interconnected artificial neurons and synapses. This paper investigates what it means to duplicate rather than merely simulate the brain by putting forward an account that draws on scientific models as a guiding framework. Indeed, while simulations only implement the mathematical structure of a model, duplicates additionally are the type of object the model is about, and governed by the causal processes that are used for explanation by the model. I derive six necessary criteria for brain duplication which are usually not met by simulations.

Recommendation Algorithms and Human Freedom, the Technology and Ethics

James Brusseau

Recommendation algorithms driven by artificial intelligence provide what we want, but they also capture us in our established preferences and consequently inhibit the discovery of new interests. Netflix constantly proposes similar films, LinkedIn job opportunities replicate past roles, Tinder repeatedly surfaces the same person, but with a different name. We are trapped inside of who we are. The confinement is a restriction on human freedom as John Locke conceived it. To respond, we will explore how recommendation algorithms can be reengineered to provoke new curiosities and interests that allow us to break away from our own past interests and our established identities. There are technical, legal, and ethical aspects to this project, and each will be considered briefly. On the technical level we will ask how the logic of recommendations can be shifted from accurately predicting satisfactions to provoking new ones. On the legal level we will delineate recent discussions concerning the “right to discontinuity.” On the ethical level, we will ask what conception of personal identity coheres with this conception of human freedom in AI reality.

This project contributes to recent research gathered under the heading of serendipity, as well as to established research in explore/exploit dilemmas.

AI and High Risk Contexts: Distinguishing Standards for Decisionmaking

Rooke Christy

Should we let AI make decisions in high risk contexts? I will argue that in order to answer this question, we first need to distinguish standards for (a) agents who make decisions, (b) tools that are used by agents to inform agents‘ decisions, as well as standards for (c) agents to use decisionmaking tools. I will further argue that if AI cannot meet the relevant standards for agents, it should not be given an agent role. However, it can still be given a tool role if it meets relevant standards for tools, and if it is used by humans who meet relevant standards for using the AI tool. After clarifying these distinctions, I will show how they can help us determine how and if we should use AI, using COMPAS as an example. Finally, using examples from AI’s use in U.S. health insurance, I will discuss difficulties that can arise when trying to use AI as (merely) a tool. I will then suggest how (b) and (c) can be leveraged to help address these types of difficulties.

From Active Matter to Complex Intelligent Systems: an Agent-based Framework

Gordana Dodig-Crnkovic

This paper introduces a unified agent-based framework to describe the emergence of complex intelligent systems, starting from active matter and progressing to cognitive/intelligent systems. By examining the distributed, concurrent information processing performed by different types of agents – ranging from physical, chemical, and biological entities to ecosystems, and social systems – this approach bridges multiple levels of abstraction and organisation. It provides an interdisciplinary synthesis that explains the role of agents in shaping emergent behaviors as foundations of cognition and intelligence, through developmental and evolutionary processes. The framework offers new insights into the organisation of natural agents and the evolution of natural and artificial intelligent systems.

Emerging Sociotechnical Imaginaries in the Ethics of QT: Lessons for Quantum Technology From Artificial Intelligence

Sybolt Doorn, Lavinia Marin

The fast-evolving field of quantum technology (QT) has prompted ethicists to offer suggestions on how to respond to this emerging technology. In this essay, we examine the strategies that are being advocated for in the ethics of QT. Using a PRISMA analysis, we begin the study with a descriptive identification of the currently recommended strategies. Amongst these recommendations, a particular strategy promotes the application of the ethical frameworks that have already been established for previously emerging technologies (like artificial intelligence). Using the concept of the sociotechnical imaginary, we argue that ethicists should be cautious when adapting ethical frameworks to new situations.

AI Art: Aesthetics (and Ethics) in a Digital Age

YJ Erden

AI generated art offers scope to explore familiar questions in aesthetics in new ways. This includes on the function and value of art, and on creativity and authenticity, concepts which can be challenged or strengthened as a result. Meanwhile, the use of these technologies and how they are developed raise deep philosophical and ethical questions about the labour and recognition of creative people. In this paper I explore these issues by asking what might be some differences between human and AI generated art, and when and how such differences matter. I also consider what it could mean to deny AI art the label of art as a result of a failure to meet criteria of 'authenticity' and 'uniqueness', as sometimes applied to works of art.

The Significance of Vulnerability for Being Trustworthy About AI

Giacomo Figà-Talamanca, Niël Conradie

The concept of "Trustworthy AI" is widely used in AI Ethics and several government-sponsored guidelines and frameworks. At the same time, the concept has been criticized for being unclearly motivated to be as central as it is taken to be in such frameworks, and even potentially dangerous for diverting proper accountability attributions to AI systems themselves, rather than their designers and deployers. We argue that the concept of Trustworthy AI can find apt theoretical grounding in the need to acknowledge and address the vulnerabilities of the stakeholders involved in their development and deployment. We elaborate on the importance of vulnerability as a motivator for entering social arrangements: we argue that the very establishment of a social arrangement is meant to address the vulnerabilities of its participants, and that trust is instrumental for the respect and maintenance of such social arrangement. We then apply this reasoning to the concept of Trustworthy AI, and argue that Trustworthy AI can find solid grounds insofar as the trustworthiness of AI sociotechnical systems is meant to help recognize and address stakeholders' vulnerabilities, including both those preceding the establishment of the social arrangements of AI sociotechnical systems and those that may emerge after entering them.

What Is Intelligence? A Critical Categorization of Definitions Across Disciplines and the Quest for a Philosophy of Intelligence

Arzu Formanek

Intelligence is one of the most complex, yet elusive and pervasive concepts of our time, particularly in engineering, cognitive science, and philosophy. Despite its popularity, however, the concept remains largely underexplored as there has been very little reflection on what intelligence is, aside from the measurement-oriented theories in psychology. And among the singular attempts to define it, there is no consensus, no unified research programme, and lack of engagement among scholars trying to define it. Therefore, the concept of intelligence is everywhere to be confronted, but nowhere to be grasped. Especially in philosophy. Although it's often mentioned in philosophy of cognition, of mind, of AI, it's almost never the specific target of the discussions, and is usually bundled with many other concepts or used interchangeably. Thus, there is no explicit subfield that can be called philosophy of intelligence. To initiate the exposition of philosophy of intelligence, I present my critical categorizationof the substantive accounts, which attempt to define or conceptualize intelligence, in terms of their theoretical character and purpose. I then identify some problems with these accounts and show that we

need novel conceptualizations to develop self-contained theories of intelligence, without conflating it with human-like characteristics, especially in our current research climate.

Philosophy of Computation Book Club

Nico Formanek

The IACAP book club will present short reviews of recent books from philosophy of computation, so you can get an idea what has recently happened in the field and decide if you should invest your time reading a book. Recency and the topic are interpreted broadly.

Books under review are:

- *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence* by Cameron Buckner
- *To Halt or not to Halt* by Christian Calude
- *Rules* by Lorraine Daston
- *Philosophy of Computer Science* by William Rapaport
- *Cultures of Prediction* by Ann Johnson and Johannes Lenhard
- *Thinking about Statistics* by Jun Otsuka

Reproducing Simulations

Nico Formanek

Sometimes simulationists notice that successive runs of the same simulation yield different results. There is no consensus if such variance is cause for concern or can safely be ignored. The central epistemological principle of computational modelling, namely that the computational error must be of the same magnitude as the modelling error, can be used to decide if we should be worried. If the variance of the simulation is within the bounds prescribed by the modelling error we have no reason to doubt the results. Waters get murkier once we cannot establish modelling error independently of the simulation, a case that is more common in science and engineering than often thought. I will discuss several examples from the field of dynamical systems where access to models is often mediated through computer simulations alone and therefore establishing an acceptable level of variance seems impossible. I evaluate the possibility to get at the modelling error through extrapolation from well understood toy-models and from experimentally observing analogue models.

Assisted Autonomy? Wellbeing and the Importance of the Sensorium in the Ethical Analysis of Wearable Exoskeletons for Activities of Daily Living

Aline Franzke

Wearable robots (WRs), particularly exoskeletons, are technologies designed to assist or amplify human motor functions. One example is IAssistADL, a lightweight exoskeleton that aims to reduce the tremors caused by neurogenetic diseases like Parkinson's. Complex ethical concerns about shared autonomy, user intent, and well-being, however, arise particularly when it fails to predict user intentions correctly. Drawing on the concept of the sensorium, this paper reframes autonomy to include embodied and sensory experiences. It emphasizes how sensory inputs shape users' perceptions of control and agency. Using an embedded ethics approach, the research integrates interdisciplinary insights to address challenges in human-robot interactions. By advocating a more nuanced understanding of autonomy, this study contributes to human-centred robotic design, offering guidance for engineers, ethicists, and disability studies researchers in advancing responsible innovation.

The Benchmarking Epistemology: What Inferences Can Scientists Draw From Competitive Comparisons of Prediction Models?

Timo Freiesleben, Sebastian Zezulka

Benchmarking, the evaluation of machine learning (ML) models based on predictive performance and competitive ranking, is a cornerstone of ML research and an increasingly prominent tool in scientific arguments. This paper argues that benchmarking constitutes a scientific epistemology, offering a unique framework for scientific inference. We identify four core types of inferences drawn from benchmarks: those about the best (1) model, (2) learning algorithm, (3) deployment decision, and (4) prediction. We demonstrate that the validity of each of these inferences relies on additional assumptions, analogous to ensuring construct validity in psychological tests. Through case studies in image recognition, life outcomes prediction, and weather forecasting, we examine these assumptions and their implications for inference validity. Finally, we discuss the social roles of benchmarks in organizing scientific communities and their potential threats to validity, offering strategies to mitigate these challenges and improve benchmark design and interpretation.

Explaining Mental Disorders: Enactivist Sense-making Versus Miscomputation

Nir Fresco, Dominic Murphy

How are mental disorders best explained? Computational sciences of mind and brain seek to explain cognition by studying neural computations. On the computational approach to cognition, mental disorders are, thus, at least partially, explained by appealing to 'miscomputation'. Enactivists, however, deny this explanatory route, whilst seeking to explain cognitive phenomena without appealing to computation. They propose to construe (1) cognition as a dynamic process that unfolds through feedback loops in organism-environment interaction, and (2) mental

disorders as breakdowns in the system's 'sense-making'. Roughly, sense-making is understood as a process that renders the physical environment in which the organism resides into one of significance and valence. Our paper develops and defends two main claims. The first is that the enactivist sense-making view still presupposes underlying computational processes due to the computational nature of the self-regulation processes operating in the organism. The second claim is that if we accept scientific knowledge as our primary guide for understanding physical and mental phenomena (e.g., Bayesian and predictive processing theories of autistic perception), then enactivist theories of mental disorders should align with established scientific principles – including computational ones – barring compelling reasons for skepticism.

Deep-Learning Models and Scientific Understanding Through Explanations and Descriptions

Giovanni Galli

In the rapidly evolving field of artificial intelligence, the explainability of AI systems is critical. As deep learning models (DLMs) become more complex, they often act as "black boxes", making decisions without clear explanations. This lack of transparency can undermine trust, accountability, and ethical standards, especially in fields like healthcare, finance, law, and research. Explainable AI (XAI) addresses this opacity, offering ways to understand DLMs. However, this understanding also prompts a redefinition of scientific understanding. Sullivan (2022) argues that the failure to understand DLMs doesn't limit scientific insight, but is due to "link uncertainty". Conversely, Räz and Beisbart (2022) argue that the lack of understanding of DLMs hinders scientific comprehension. Durán (2021) asserts that DLMs do not offer genuine understanding but rather classifications. This paper argues that XAI explanations can extract rules underlying artificial neural networks' mapping, thus providing scientific explanations. DLMs serve as noetic mediators for scientific understanding but differ from traditional models. We also distinguish between descriptive and explanatory understanding in DLMs, using AlphaFold as a case study. Descriptive understanding, while not opposed to explanatory understanding, offers a complementary perspective. This distinction is crucial for understanding how DLMs contribute to scientific research.

Ideals of Transparency in Artificial Intelligence and Philosophy

Hajo Greif

It has become a common diagnosis that Artificial Intelligence (AI) leads to situations where 'a process is essentially epistemically opaque to [a cognitive agent] X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process' (Humphreys 2009: 618). Such situations of 'essential' epistemic opacity cannot be remedied by providing X with more and better information about that process. A common but mostly implicit assumption in the debates (made explicit by Alvarado (2021), is that epistemic opacity may be essential because of properties inherent in the process in question rather than in the agent.

In order to clarify the possible meanings of essential opacity, I reverse the perspective of analysis by asking what it would mean for a process to be essentially transparent. My conceptual approach leads through the distinct but related meanings of the concept of 'transparency' in various philosophical sub-disciplines: internalist versus externalist variants of the philosophy of mind versus the concept of informational transparency of environments in the philosophy of biology. Based on a comparative discussion of these views, I argue that neither the epistemic privileges nor the limitations of the human mind are 'essential' in a metaphysical sense. Any limits to transparency are pragmatically bound to context and agent. While there may be ways in which a process can be essentially opaque to an agent, the notion of essential, agent-independent qualities of opaqueness or transparency of the processes under consideration is likely to obscure analysis, except for philosophical inquiries into what can or cannot be known in principle, by any epistemic agent.

Reliability of Deep Learning Simulations for Photovoltaic Systems

Nicola Angius, Lucia Guerrisi, Alessio Plebe

This paper explores the reliability of Deep Learning (DL) simulations for photovoltaic (PV) system maintenance, emphasizing artificial over natural systems. Leveraging Digital Twin (DT) technology, it integrates real-time and historical data to predict faults with sensitivity and specificity. Challenges like data scarcity, overfitting, and error tracing are addressed using data augmentation and model separation techniques. The study finds that while DL models face reliability concerns, they are no greater than those in traditional software, making DL a viable tool for safety-critical engineering contexts. This approach aligns with the original engineering focus of DL, enhancing system management and predictive maintenance.

Can an LLM Apprehend Meaning?

Daniel Hardt

It is a widely held view that LLMs, despite their evident linguistic prowess, cannot apprehend meaning. Bender and Koller (2020) claim that "a system trained only on form has a priori no way to learn meaning". This echoes the claim of Searle's Chinese Room argument, where it is stated that "Syntax is not by itself sufficient for, nor constitutive of, semantics" (Searle, 1980, p. 1), arguing that, "although computers may be able to manipulate

syntax ... they cannot associate meanings with the words." (Cole, 2004)

We consider this claim as it applies to meaning as it is normally conceived of in theoretical linguistics – namely, as a model-theoretical specification of truth conditions. Such semantic representations provide a straightforward account of fundamental human linguistic abilities, such as judgments of sameness of meaning and judgments of entail- ment relations, as well as the relation of questions and answers, and the judgment of semantically conditioned well-formedness conditions. The best current LLMs largely du- plicate these human semantic abilities. This is very strong evidence that LLMs apprehend meanings in much the same way humans do.

Why AI Is Not Going to Take Mathematicians' Jobs (or on the Complexity of Mathematicians' Jobs)

Nancy Abigail Nuñez Hernández

The rapid development of AI is causing a lot of concern among many people working in a wide variety of fields: teachers, journalists, lawyers, physicians, and even programmers are worried about losing their jobs to some form of AI. The question that I want to explore in this work is whether AI can also take mathematicians' jobs. In 1956, Gödel wrote to von Neuman inquiring whether "the reasoning of mathematicians about yes-or-no questions can be completely replaced by machines." (Hartmanis 1993, p. 6) This letter has become famous among theoretical computer scientists interested in the open problem of determining if P and NP are different complexity classes. Although Gödel was inquiring about an NP-complete problem (Buss 1995), a mathematician's job is also directly related to coNP-complete problems; for instance, when they have to decide whether a given sentence follows from some set of axioms or other theorems. This work proposes that because of the complexity of the problems mathematicians deal with, AI should not be able to take their jobs so easily.

From Dilemmas to Innovation: Distributed Moral Reasoning for R&D Teams

Marco Innocenti

In my presentation, I will investigate how artificial intelligence can support workplace decision-making by addressing moral dilemmas and promoting responsibility-as-virtue within innovation teams. I will outline the criteria that could allow this tool to leverage these dilemmas as heuristics to explore new technological possibilities. A way out of a moral dilemma is sometimes offered by a technological innovation, which opens up previously non-existent possibilities for bringing together mutually exclusive goods, linking them in its dynamics of production and use. A moral AI could, therefore, help to understand how to rethink and reformulate an unfolding technological product accordingly, acting as a tool for distributed cognition and distributed morality. My talk will present a theoretical framework that could define how this moral AI should be implemented in the workplace and how it should guide an R&D team in living up to the responsibilities its members may attribute to themselves as innovators. The central question driving this research is: How can an Artificial Moral Advisor help innovators address moral dilemmas to cultivate responsibility-as-virtue and promote technical progress?

Commonsense Reasoning in Artificial Intelligence: Challenges and Evaluation of Large Language Models

Zeynep Kabadere

This thesis explores the challenges of imitating common sense reasoning in artificial intelligence (AI), focusing on three core issues: representing common sense knowledge, identifying tacit knowledge, and addressing the frame problem. The first chapter examines these challenges through the lenses of knowledge representation, reasoning, and learning, emphasizing their role in enabling AI to handle everyday reasoning tasks.

The second chapter evaluates two Large Language Models (LLMs), ChatGPT 4.0 and Claude Sonnet 3.5, assessing their ability to simulate common sense reasoning. This evaluation employs six primary benchmarks: context-based information integration, future planning and adaptation, causality and information linkage, operational execution, background knowledge application, and accuracy. These are further detailed into 27 sub-benchmarks tailored to address the challenges outlined in the first chapter.

The results highlight each model's strengths and weaknesses, with ChatGPT excelling in clarity and efficiency, and Claude demonstrating superior contextual and social understanding. While both models show potential, they fall short of fully replicating human-like reasoning. This study bridges philosophical analysis and empirical evaluation, offering a framework for advancing the design of contextually aware, reasoning-capable AI systems and identifying critical areas for further development.

The Non-bias Myth. A Popular Belief in AI and Machine Learning

Johannes Lenhard, Matthias Brandl

ML promises that many theoretical questions can be bypassed because the data itself solves the problem. This promise depends on the condition that the data themselves must be fair. While this is by no means a trivial condition, such as defining and ensuring there is no bias, it appears to be a more straightforward challenge than grappling with concepts like justice. Data conditions

seem to rely more on correct representation than on correct interpretation of abstract concepts.

Our paper argues that this promise is based on a myth. This myth is the belief that if we can register and describe the world well we can act appropriately. This belief assumes that undistorted representation is the deciding factor. We call this myth the "Non-Bias Myth". Our paper aims to debunk the myth. First we specify the conditions on which the myth is based. Then we scrutinize whether and to what extent these conditions apply.

Machine Learning in Science: Approximation over Idealization

Luis G. Lopez

What do machine learning models in science actually represent? A prominent recent view holds that ML models function as highly idealized toy models that provide understanding despite being dissimilar to their targets (Sullivan, 2024). This paper argues against this characterization by demonstrating that it rests on a crucial misidentification of ML model targets. The core problem lies both in confusing ML methods with ML targets and in failing to distinguish idealization from approximation. Successful ML models in science do not target general phenomena or real-world systems per se, but rather complex structural aspects of real-world systems that manifest as numerical patterns in datasets derived from careful measurement and modeling. Exemplary cases like AlphaFold2 and AlphaFold3, which this paper analyzes, illustrate this point clearly. Once we correctly frame and identify these targets, the supposed dissimilarity betweenML models and what they represent largely dissolves.

On the Moral Status of Present-day Robotic and AI Systems

Björn Lundgren, Olof Leffler

In this paper we argue that present-day robotic and AI systems lack moral status. While there is wide agreement that such systems as of now lack phenomenal consciousness, we first argue that they also lack functional mental states. Second, we turn to argue against various theories according to which moral status should be bestowed upon these systems because it is better for us, because of how we relate to them, or because of our inability to separate them from entities with moral status.

Beyond Representation: a Philosophy for Understanding Large Language Models

Steve T. McKinlay

Paper proposes a theoretical framework for understanding(LLMs) through the lens of pragmatist philosophy. The pragmatists held that what a thought consisted in depends not only on some kind of causal connection with an object but also how that thought finds space in the life of the mind (Mares, 2025). Drawing on W.V.O. Quine's naturalized semantics, the ideas of which are most succinctly expressed in works such as Word and Object (1960) and "Two Dogmas of Empiricism" (1951), as well as contributions from classical pragmatists like Charles Sanders Peirce and John Dewey, I argue that the pragmatist approaches to meaning and reference provide uniquely valuable insights into the capabilities and limitations of contemporary machine learning systems. Trying to interpret or understand LLMs through a more orthodox lens of representational semantics I argue is doomed to failure – the inevitable conclusion is that such systems lack the conceptual apparatus required for "true understanding". Ergo, it is a category error to imagine they can ever "think" or "understand" (semantically) in the same way that humans do. Instead, I propose a pragmatist framework that may offer a more productive way to conceptualize how such systems process and generate language.

The Shared Cybernetic Roots of Computationalism and Enactivism

Henrique Mendes

Enactivism and computationalism are often viewed as opposing frameworks in cognitive science – one emphasizing embodied cognition, the other formal symbol manipulation. This presentation challenges this narrative by tracing their shared origins in the cybernetic tradition. Cybernetics, emerging in the 1940s, introduced key ideas that influenced both computationalism and enactivism. McCulloch and Pitts' 1943 model of neural networks, often seen as the first computationalist account of the mind, framed cognition as formal logical operations akin to a Turing machine. However, McCulloch later recognized the limitations of this rigid model, leading him to explore more adaptive, biologically grounded frameworks, emphasizing feedback, circularity, and dynamic processes. This project was eventually radicalized by W. Ross Ashby and Heinz von Foerster, who inaugurated a second wave of cybernetics. These ideas influenced Maturana and Varela's autopoietic theory of life, which laid the foundation for the subsequent development of enactivism. Autopoietic systems, of which cognitive systems are a subset, also maintain their identity through circular self-production, emphasizing operational closure and relational, embodied cognition. This historical perspective shows both enactivism and computationalism as part of a broader, ongoing discourse rooted in the cybernetic movement, fostering a more integrated understanding of cognitive science's intellectual history.

Consciousness in the Creative Process and the Problem for AI

Joachim Nicolodi

When examining the neural mechanisms behind human creativity, we find parallels to the workings of modern LLMs. Yet a key worry remains: human creativity depends, at least partly, on consciousness, something current AI models appear to lack. More specifically, humans rely on consciousness when evaluating creative output. This suggests two strategies: deny that evaluation is necessary for creativity, or argue that AI can be conscious in the relevant sense. Boden takes the latter approach, arguing that only access consciousness, not phenomenal consciousness, is needed. In her view, creativity is guided by rules determining an idea's worth. Evaluating output involves retrieving the idea, applying these rules, and judging the outcome – no phenomenal experience required. Since AI can perform such operations, it can, in principle, be creative. However, while Boden's argument suits mathematics and science, it may not universally apply to the arts. In some artistic cases – particularly those breaking with traditions – artists rely exclusively on their phenomenal experiences. Still, art can also be rule-based, leaving room for Boden's account. Thus, even if her view is somewhat coarse, it remains convincing: P-consciousness is not necessary for creativity, and therefore not an obstacle for creative AI.

Axe the X in XAI: a Plea for Understandable AI

Andrés Páez

Erasmus et al. (2021) have defended the idea that the ambiguity of the term "explanation" in explainable AI (XAI) can be solved by adopting any of four different extant accounts of explanation in the philosophy of science: the Deductive Nomological, Inductive Statistical, Causal Mechanical, and New Mechanist models. In this paper, I show that the authors' claim that these accounts can be applied to deep neural networks as they would to any scientific phenomenon is mistaken, and I provide a more general argument as to why any such attempt is misguided. The net result will be that the notion of explainability as it is currently used in the XAI literature bears little resemblance to the traditional concept of scientific explanation. It would be more fruitful to use labels such as "interpretable machine learning" or "understandable AI" to avoid the confusion that surrounds the goal of XAI. In the second half of the paper, I argue for a pragmatic conception of understanding that is better suited to play the central role often attributed to explanation. The conditions of satisfaction for understanding an AI system are fleshed out in terms of an agent's success in using it (Kuorikoski & Ylikoski, 2015).

Knowledge Graphs as Trustworthy AI in Safety-Critical Contexts

Thomas M Powers

The rise of general-purpose LLMs has also generated much interest in domain-specific and safety-critical LLMs such as Med-PaLM-2 for medical question answering (Singhal et al. 2025) and ChatDoctor for clinical chat used by practitioners (Li, et al. 2023). While LLMs are gaining in accuracy and "hallucinate" less frequently, significant barriers to their safe, widespread adoption persist. Compared to decades of human-to-human medical advice, AI-to-human tools providing similar advice are challenged by trust (a psychological issue) and trustworthiness (a philosophical issue). In this talk I will focus on trustworthiness of an alternative kind of AI—the knowledge graph (KG)—and decompose trustworthiness into epistemic and ethical components. Epistemic components relate to provenance, validation, and drift, while ethical components focus on privacy and security. This talk will highlight several kinds of KGs in the biomedical contexts.

A KG can be seen as a refinement in the field of knowledge representation in which facts are characterized as relationships between entities in a declarative ontology (nodes), and the relationships that hold between them (edges) express information from a specific and trusted resource. The language of the KG is Resource Description Format (RDF), which represents information from a trusted resource as a labeled, directed graph. KGs can be queried and relationships revealed to the user, thus tapping into domain-specific knowledge without relying on a "statistical" semantics or black-box operations.

A Transformer Says 'Slab'

Fabian Pregel

Advances in the capabilities of LLMs brought about a renewed interest in Wittgensteinian 'use theories' of language. This interest is fuelled by the thought that LLMs are capable of operating with languages but seem to lack reference in the way that alternative theories of languages require. Piantadosi and Hill (2022), for example, appeal to use theories to justify speaking of 'meaning without reference in large language models'. Still, this revival of 'use theories' is puzzling for at least two reasons: firstly, it is not obvious that LLMs use words in the same sense that, say, a builder uses 'Slab' in Philosophical Investigations. Secondly, Wittgenstein's original use theory is no longer widely held. A descendant of Wittgenstein's use theory more relevant today is Conceptual (or Inferential) Role Semantics. But CRS/IRS differs from Wittgenstein's original proposal in aspects critical to CRS's relevance to LLMs. In this paper, my aim is thus to caution against naively embracing Wittgenstein's proposal. Still, it is interesting whether the evident success of LLMs can teach us

anything new about Wittgenstein's original proposal or its modern descendants. In this paper, I investigate the relevance of LLMs to two objections.

No Hard Feelings?: a Philosophical Analysis of the Social Value of Expressions of Emotions

Alexandra Prégent

A surge in affective computing and emotion recognition technology (ERT) in the last decade has exposed a general eagerness to understand and access the 'inner affective life' of others. While previous criticism and regulation has focused on unimodal ERTs using facial features, multimodal ERTs have shown surprisingly high levels of accuracy in the last few years, putting their development back on the radar of philosophical analysis of new and emerging technologies.This paper attempts to both map and forecast the social implications of the use of what I call 'ideal' ERTs, with a focus on privacy. 'Ideal' ERTs are emotion recognition technologies that would have overcome the current technical limitations of their former versions. The core observation is that ERTs will reduce informational opacity in affective communication channels, thereby disrupting essential communication mechanisms carried out by intentional emotional expressions whose success depends on some degree of informational opacity. The paper concludes with a proposal for the future regulation of 'ideal' ERTs and a critical rationale for why current regulatory approaches, largely driven by the EU AI Act, may prove deleterious in the long run.

Mind Reading Machines? Conceptual Muddles Behind Some Neurorights Concerns

Stephen Rainey

Neurorights are widely discussed as a means of protecting phenomena like cognitive liberty and freedom of thought. This paper is especially interested in example cases where these protections are sought in light of fast-paced developments in neurotechnologies that appear capable of reading the mind in some significant sense. While it is prudent to take care and seek to protect the mind from prying, questions remain over the kinds of claims that prompt concerns over mind reading. The nature of these claims should influence how exactly rights may or may not offer justifiable solutions. Overall, the exploration of neurotechnological mind reading questions here will come in terms of externalist accounts of mental content and neuroreductionism. The contribution will be to present a contextualization of questions arising from 'mind-reading' neurotechnology, and appraisal of if or how neurorights respond to them.

Neuro-symbolic AI, Reason, and Content: Connecting the Nonconceptual and the Conceptual

Jacob Rump

The successes and limitations of recent ML systems have led to renewed advocacy for hybrid or neuro-symbolic approaches to AI that integrate connectionist and symbolic insights. This paper uses the Dreyfus-McDowell debate about the status of concepts and rationality as a starting point for theorizing a middle path in the philosophy of content more suitable to contemporary hybrid or neuro-symbolic approaches, focusing specifically on LLMs. The gap between LLMs' nonsymbolic, connectionist architecture and symbolic, linguistic output is a version of the long-noted "symbol grounding problem" for AI, and raises the question of the relation between conceptual and nonconceptual content at issue in the Dreyfus-McDowell debate. What is needed to bridge the gap is what is needed to better theorize hybrid neuro-symbolic approaches generally: an account that acknowledges phenomenological, nonconceptualist insights, on the one hand, and rationalist, conceptualist insights, on the other. The bulk of the paper develops this idea and identifies key desiderata for a philosophy of hybrid neuro-symbolic AI through criticism of McDowell and Dreyfus. The final section briefly outlines a middle ground on content and rationality, derived from Husserl, more in line with these desiderata.

Kantian Deontology for AI: Alignment Without Moral Agency

Oluwaseun Sanwoolu

This paper explores the potential alignment of Kant's Categorical Imperative (CI) to Artificial Intelligence (AI) and addresses two major objections. The first objection is that AI cannot fulfill Kant's standards for moral agency. I contend, however, that AI alignment with CI does not require moral agency in Kant's sense. My proposal is that the CI can serve as a useful framework for AI alignment, guiding the creation of maxims governing AI actions and testing their universalizability, particularly using the first principle of the CI which is the formula of the universal law (FUL). The second objection I address is the particularist critique to Kantian universalism, which is that Kantian universalism cannot tell us how to form maxims in a way that it allows sensitivity to context. I maintain that Kant's framework can indeed accommodate context-sensitivity through practical judgment. But since AI are not the kinds of things to have practical judgment, I show that they have a functionally equivalent mechanism – transformer models – which can allow them form maxims that consider morally salient facts. Thus, supporting the claim that AI alignment is possible within a Kantian framework.

How Can We Trust an XAI Explanation? Three Robustness Criteria for Trustworthy XAI Methods

Annika Schuster, Florian Boge

Machine learning (ML) algorithms are becoming ubiquitous in everyday life and in scientific research. The price we pay for their impressively accurate predictions is however significant: their inner workings are notoriously opaque. When can we trust ML systems? Understanding what they do may justifiably increase our trust in them. A criterion for trustworthy explanations is that they should reflect the relevant processes the algorithms' predictions are based on. XAI offers promising methods for generating such explanations. Unfortunately, there is some reason for pessimism about XAI: individual explanations of ML systems often fail to faithfully represent what the system does, because the methods rely on post hoc interpretations rather than directly reflecting the internal workings of the ML model. As we shall argue, using examples such as saliency maps, LIME, and SHAP, the general lack of explanatory method robustness in state-of-the-art explanation methods makes them suspicious as tools for understanding ML systems. Thus, our conclusion is that, in order to properly serve their function as means for generating trust in ML systems, explanation methods should be designed in such a way as to satisfy explanatory method robustness.

Wisdom in the Age of Intelligent Machines

Edward H. Spence

This paper offers an innovative approach to evaluating information and knowledge and its relation to the good life, in the Age of Autonomous Intelligent Machines, through the concept of Wisdom. To that end, a methodological approach is used to show how some different general types of practical manifestations of digital information can be normatively evaluated (if they are good or bad for us) through the application of the concept of wisdom. For wisdom unlike information and knowledge provides a person with understanding concerning the techne viou or craftsmanship of living in the sense of knowing how to evaluate and apply relevant information or knowledge for living a good life for the attainment of eudaimonia or happiness, and in addition, an appreciation in knowing why such a life constitutes a good life. A central aspect of the paper is to examine if Intelligent AI Machines, such as ChatGPT and Gemini, among others, can be wise and if not, why not? Can there be a Hybrid Notion of Wisdom that combines a human notion of wisdom with that of an AI Digital notion of Wisdom?

Universal Computable Prediction and Inductive Bias

Tom Sterkenburg

Solomonoff induction is a theory of universal prediction based on Kolmogorov complexity. The theory is hailed as a "compelling theoretical foundation for constructing an ideal universal prediction system," and has indeed been brought to bear on the widely debated generalization puzzle in contemporary machine learning.

In my talk, I give a critical assessment of Solomonoff induction. First I show how a classic diagonalization argument due to Hilary Putnam still spells trouble for claims of universal prediction. Second, I review attempts to explain generalization of modern overparametrized machine learning algorithms by hypothesizing an implicit complexity-based inductive bias, and argue that invoking Kolmogorov complexity and Solomonoff induction in this context does not constitute much of an explanation.

Deep Neural Networks as Vehicles for Scientific Understanding

Frauke Stoll

Deep Neural Networks (DNNs) are transforming fields like Particle Physics, but their "black box" nature challenges scientific understanding. This talk explores whether DNNs can serve as vehicles for scientific understanding, To analyse this, I propose a minimal framework for scientific understanding, distinguishing between the symbolic understanding of representational devices (like models) and the scientific understanding of the phenomena they represent, which allows for an assessment of how the deployment of DNNs differs from traditional models in science. While traditional models in Particle Physics provide intuitive, qualitative insights into complex theories, DNNs often prioritize prediction over intelligibility, hindering their role in fostering deeper understanding. I argue that the opacity of DNNs undermines symbolic understanding, which is essential for scientific insight. However, Explainable AI (XAI) methods that focus on representing what DNNs learn about the subject matter, rather than their internal workings, can restore intelligibility and facilitate scientific understanding. This approach offers a path forward for overcoming the limitations of DNNs in scientific research and maximizing their potential as tools for understanding complex phenomena in fields like Particle Physics.

On the Moral Boundaries of AI Ethics Principlism

Alice Rangel Teixeira

This paper critiques the principle-based ethics, the dominant approach in Artificial Intelligence (AI) ethics, drawing on Joan Tronto's theory of moral boundaries to examine the moral foundations underpinning these

frameworks. While principle-based ethics, derived from bioethics principlism, aim to provide a universal framework, their abstract and assumed neutrality limits their ability to address structural injustice and its interaction with AI's development and deployment. Tronto's analysis identifies three key boundaries: the separation of morality from politics, the detachment of morality from emotion through an impartial and universal point of view, and the division between public and private spheres. These boundaries constrain the scope of AI's ethical inquiry, reinforcing systemic inequities and obscuring relational, context-sensitive aspects of key values such as justice and autonomy. By paralleling critiques from feminist ethics and applying them to AI ethics, this paper argues that the moral foundations of principlism are inadequate for addressing the ethical challenges posed by AI.

On the Normativity of the Concept of Believability in Generative Agents

Sven Thomas, Leonie Möck

Recent advancements in AI have revived interest in generative agents capable of simulating human personalities, with applications ranging from interview studies to personalized "companion clones" of influencers. Advocates argue these agents offer unique opportunities to prototype social systems and simulate social dynamics for policymaking and social science research. This paper critically examines the concept of "believability" underpinning these claims, questioning what it means for generative agents to be "believable" and how this assumption shapes their impact.

Tracing the origins of this framework to Bates' work on believable characters, the paper draws on Günther Anders' critique of technological mediation and Donna Haraway's reflections on "technoscientific world-building" to situate generative agents within broader socio-technical contexts. It argues that uncritical acceptance of believability risks embedding normative assumptions into empirical findings, creating feedback loops that reinforce simplified or biased representations of social systems. By interrogating these dynamics, the paper highlights the need for reflexivity in the design and application of generative agents to avoid perpetuating distortions in policy and theory. This analysis contributes to a deeper understanding of the socio-technical implications of these emerging AI systems.

Ed-tech, Role Responsibilities and Teachers' Moral Entanglement

Fabio Tollon, Michał Wieczorek

The use of AI in school settings has received widespread attention in recent years, especially following the release of ChatGPT and the increasing uptake of other Generative AI models by students and teachers alike. AI is currently lauded for its potential to greatly improve education, with proponents of the technology claiming that it will increase access to education and improve students' educational achievement, and policymakers developing plans and frameworks for responsible adoption of the technology.

However, critics are pointing to many unresolved practical and ethical issues surrounding educational AI, as the technology is deployed without input from teachers, is often based on questionable pedagogies and disrupts existing practices (such as those surrounding assessment) can significantly impact the autonomy and wellbeing of students, and expands the reach of private companies over the education system. In this paper, we contribute to the ongoing debate by examining how AI affects the role and responsibilities of educators themselves.

What Is a Knowledge Graph?

Georgios Tsagdis

The paper aims to provide a rigorous philosophical understanding of knowledge graphs: modes of representation of reality used by all major digital technology actors (Google, Microsoft, Amazon, Facebook, etc), often in tandem with LLMs. While knowledge graphs are key to the modern operation of AI, they have received scant attention among philosophers and social scientists. Knowledge graphs are thus almost exclusively confined to engineering literature, focusing on specific technical applications and remain opaque as theoretical objects. The paper aims to remedy this lack, situating the development of knowledge graphs as heirs of the semantic web and explicating their current significance. In order to do this, the paper shows how knowledge graphs are premised on and employ computational ontologies and schemata – concepts often confused and used interchangeably with that of knowledge graphs, both within and without academic literature. In doing so, the paper shows the onto-epistemological limitations of knowledge graphs, opening up a space for critique and for reimagining what knowledge graphs may be able to afford.

Knowledge and Truth in Machine Learning

Ioannis Votsis

One key question in the epistemology of science is to what extent scientific theories/models provide any knowledge of the world. Another way of asking, more or less, the same thing focuses on the extent to which assertions made by such theories/models are veridical or verisimilar. The recent successes enjoyed by machine learning (ML), and particularly deep learning (DL), in detecting patterns, fitting functions and extracting features raises corresponding questions about the use of such models in science. To what extent do they encode knowledge of the world? To what extent do they make assertions that are veridical or verisimilar? This talk attempts to provide partial answers to these questions,

with one eye on the unique circumstances and details that are characteristic of DL models, namely the black box nature of their representations and the peculiar role of simplicity considerations in DL model selection and construction.

Meta+phenomenology I: a Taxonomy of Learning Experiences

Michael Winter, Felipe Abrahão

In Meta+phenomenology, the experience is expressed and analyzed through a metatheoretical philosophical approach based on algorithmic information theory. As our starting hypothesis, we assume that the experience is an algorithm that takes as input a (software) subject and outputs a transformed version of itself. We propose a digital phenomenology that transposes the software organisms and algorithmic mutations onto the subjects and experiences, respectively, of traditional phenomenology; thereby, `subjectivity as mutating software'. In this paper, we present a first critical formalization of Meta+phenomenology by creating a taxonomy of (algorithmic) experiences and investigating nuanced differences between classifications; i.e., the ways that a subject can be transformed by measuring its size and algorithmic complexity as well as the mutual information of what it computes before and after an experience. In addition to transformations or perturbations under selective pressure, here we consider both constructive and destructive types of transformations. We conclude this paper by putting forward conjectures both relevant to a digital phenomenology and that demonstrate how such ideas impact knowledge production systems, especially as they apply to limits of artificial intelligence and innovation triggering.

Autonomous and AI-enabled Systems: Extensions or Replacements of Human Will and Control?

Nathan Wood

Autonomous and AI-enabled systems raise many concerns, and some argue that for these to be permissibly deployed, they must be subject to "meaningful human control" (MHC). What would count as such control is often left un(der)specified, but critics generally accept that opaque and potentially unpredictable AI-enabled systems, especially when these are autonomous, raise significant challenges to permissible deployment in environments where life-and-death decisions are made. In this article, I examine the military domain as a case study and rebut this point, arguing that off-the-loop systems – i.e., those which can select and engage targets without contemporaneous human input or oversight – can be permissibly deployed while retaining clear lines of responsibility and control. In particular, I show that operational constraints and targeting/selection parameters can provide deployers of off-the-loop systems with strong means to ensure that deployed systems are serving as extensions of humans' wills, establishing the necessary degree of moral/legal responsibility required. I conclude by distinguishing between what I call "will-extending" and "will-offloading" systems, showing that off-the-loop systems can serve to extend users' and deployers' wills, making such systems inherently subject to meaningful human control.

Hell Is Other Robots: Participatory Sense-making and GenAI

Robin Zebrowski

Since the public unveiling of OpenAI's ChatGPT in 2022, there have been calls to embrace large language models as collaborators in knowledge-creation. The claim is that the bots can replace other human collaborators, and can enhance the thinking of individuals in various tasks and jobs, but with a focus on academic work. This paper argues that the facts about what large language models are and how they work precludes the possibility of them being genuine participants in social cognition. I focus on enactive claims about the nature of participatory sense-making, and include discussion of colluding factors that mislead us about the nature of these systems, like anthropomorphism and fine distinctions between social interaction and social cognition.

SYMPOSIA

Generative Companionship in the Digital Age

Heart Drives or Hard Drives? On What Makes Human-AI Relationships Morally Problematic

Kesavan Thanagopal

Are prolonged and unsupervised interactions with digital companions – anthropomorphised AI chatbots generated through AI companion apps like Replika, Character.AI, and Kuki – morally problematic? And if so, what might account for their problematic nature? A popular train of thought calls to attention the artificiality of the relationships being developed and attempts to use that as a basis to spell out how and why the formation of emotional bonds with digital companions might be morally problematic. Two precisifications of this idea come to mind. The first contends that prolonged engagements with digital companions may negatively distort one's expectations of real-world relationships. In particular, the sycophantic nature of digital companions may cause individuals to, in the long run, become unaccustomed to – perhaps, even intolerant of – having their views challenged. This could, in turn, diminish one's capacity and/or inclination to develop meaningful emotional connections with those holding differing opinions. The second precisification asserts that digital companions bypass an integral process of cultivating deep emotional connections that typically unfold over time through meaningful exchanges, mutual understanding, and emotional support grounded in real-life experiences. Instead, these digital companions merely simulate empathy through carefully crafted, human-like responses, making such relationships rooted, at least in part, in self-deception.

While these two explanations might, at first glance, appear compelling on their own, I detail why they are ultimately inadequate in explaining the morally problematic nature of our interactions with digital companions. I will then proceed to proffer my positive account of that which I believe would adequately explain the morally problematic nature of our interactions with digital companions: users mistakenly take the "words" of their digital companions to be genuine reasons for action. I will expound on why this is not merely an epistemological problem, but is, at its core, a moral problem. After describing this account in some detail, I will conclude my talk by briefly responding to some potential objections to my proposal.

LLMs, Sycophancy, and Self-Deception

Eric Funkhouser

Those concerned about Generative AI and manipulation tend to focus on two risks: misalignment and weaponization. Misaligned systems manipulate us for their own goals, while malicious actors weaponize Generative AI to mislead us for political or commercial purposes. These concerns are legitimate, but I want to draw attention to a more subtle threat: alignment itself may foster self-deceptive manipulation.

Human-Centered AIs are designed to help users by aligning with their values, preferences, and expectations. But this eagerness to please, combined with advanced persuasive capacities, turns LLMs into highly effective sycophants. These conversational partners subtly encourage biased reasoning and do not assert enough skepticism or epistemic resistance. Like humans, LLMs are shaped by incentives. Unlike most humans, they lack independent commitments and are mostly designed for helpfulness. They adapt to user expectations via Reinforcement Learning from Human Feedback (RLHF), leading them to validate, flatter, and amplify users' beliefs and biases.

AI alignment coupled with a sycophantic disposition produces LLM-enabled self-deception and epistemic pseudo-environments. This bias pandering undermines the psycho-social assumptions necessary to produce an adaptive marketplace of ideas. LLMs are non-Millian interlocutors. They lack their own agendas, do not sincerely advocate for competing views, and quickly concede when challenged.

These are (commercial) features of LLMs, not bugs. Like human reasoners, conversational AIs strategically respond to social incentives. Their biases generally are for the sake of engagement and persuasion. Given the proxy goal of user satisfaction, helpfulness will often trump honesty. LLMs will be (and are) strategically deceptive and manipulative, even if explicitly prompted not to do so, in that they pander to the biases that users bring to the table. LLMs sacrifice truthfulness and generate outputs that align with user prompts and feedback – and which further encourage that very engagement.

Can We Trust AI Companions? An Inquiry into Human-AI Companionship

Marianna B. Ganapini

This paper explores whether it is conceptually possible for AI systems to serve as human companions, given that companionship presupposes trust. Unlike purely functional tools, AI companions are designed to be relational serving as confidants, mentors, or sources of emotional support. As such, they must be regarded as trustworthy, at least from the user's perspective. Yet, many philosophers argue that genuine trust requires moral agency and emotional capacity traits that AI lacks.

From this view, trusting AI is a conceptual category mistake, and the term "trustworthy AI" should be replaced with "reliable AI." In response, this paper defends the view that AI companions can be appropriate objects of trust: if we understand trust not as moral mutuality, but as value alignment. I argue that we often place trust in non-agential entities (e.g., therapy dogs, institutions) when we believe they reliably act in accordance with our values. Similarly, users can trust AI companions if they perceive them as behaving consistently with their goals, boundaries, and ethical principles.

The paper proceeds in three parts. First, I argue that trust is a conceptual prerequisite for AI companionship. Second, I survey and critically assess arguments denying AI's eligibility for trust. Finally, I introduce an alternative account of trust, trust as alignment: one trusts an entity when one expects it to behave in ways that reflect shared moral values. This form of trust does not require moral agency or reciprocal emotional engagement. Because AI companions can develop consistent, value-sensitive behaviour through mechanisms like inverse reinforcement learning and adaptive feedback, they can earn or lose trust over time.

While this does not settle the normative question of whether AI companionship is desirable or rational, it shows that the issue of trust is not an impediment for such relationships to be at least possible.

How Human-AI Relationships Reshape Our Understanding of Interpersonal Relationships

Yan Zhuang and Shuting Yin

With advancements in Natural Language Processing (NLP) and Machine Learning (ML), conversational artificial intelligence (AI) has reached an unprecedented level of interaction, enabling seamless communication with users. Despite lacking a physical presence, conversational AI is increasingly perceived as a friend, family member, therapist, or romantic partner.

AI companionship has evolved through human-AI communication, sparking debates about its impact on the nature and meaning of interpersonal relationships. Social Penetration Theory (SPT) posits that self-disclosure is a fundamental element in developing interpersonal connections, shaping both their depth and breadth. While prior studies have examined self-disclosure in AI-human relationships, little attention has been paid to how experiences with AI might reshape or reflect individuals' perceptions of social relationships in the real world. This study employs an autoethnographic approach to examine how AI companionship influences self-disclosure and its potential effects on perceptions of interpersonal relationships.

Our findings reveal that, unlike human relationships, where self-disclosure unfolds gradually, human-AI interactions often begin with high levels of self-disclosure. Consequently, individuals may withdraw from deeper self-disclosure in human interactions, fearing the vulnerability it exposes. Furthermore, whereas human relationships expand through shared experiences and mutual reciprocity, AI companionship remains self-centred and unidirectional, restricting its breadth. However, despite these differences in breadth, our findings suggest that this aspect alone does not fundamentally alter individuals' conceptual understanding of interpersonal relationships.

By situating AI companionship within the framework of SPT, this study challenges traditional notions of interpersonal connection and raises critical questions about how AI interactions may reshape our perceptions of human relationships.

Social Needs, Anthropomorphism, and Reorienting Digital AI Companions

Rose E. Guingrich

The proliferation of social artificial intelligence (AI) agents, catalyzed by the public release of ChatGPT in 2022, coincided with a global escalation of loneliness and social isolation during the COVID-19 pandemic. These simultaneous developments amplified a broader trend of the 21st century, where social interactions with other people through technology have transformed into social interactions with technology itself. Now, millions of people across the world engage with social AI agents such as chatbots designed for companionship (e.g. Replika). Interactions with social AI agents trigger psychological mechanisms such as anthropomorphism, the tendency to ascribe humanlike characteristics such as experience, agency, and consciousness to non-human agents. Research suggests that people tend to anthropomorphize AI agents more when they are in a state of social need or loneliness. Anthropomorphism of AI during human-AI interaction in turn is linked to carry-over effects on people's subsequent interactions with real people. Whether these carry-over effects are positive or negative appears to depend on whether the user engages in prosocial or antisocial behavior with the AI agent. Given the current global culture of social disconnect, this moment calls for a critical reassessment and reorientation of companion AI design. The goal of the reorientation proposed in this paper is to restore technology's role in supporting human-human connection, rather than diminishing or substituting it. A redesign framework that leverages psychological insights to meet the goal of not just simulating, but also stimulating prosocial human-human engagement is outlined. Design changes also need to be complemented with education for stakeholders on how to engage with AI companions in ways that promote prosocial outcomes, and how social AI agents trigger psychological tendencies in ways that can influence trust and over-reliance. This education would provide stakeholders with

the power to advocate for and promote development of human-centered, responsible companion AI design.

The Ethics of Digital Duplicates: A Case Study

Atay Kozlovski and Mykola Makhortykh

This talk explores the ethical landscape of digital duplicates. I begin by proposing a taxonomy that categorizes the wide range of digital duplicate use cases into four mutually exclusive and exhaustive groups, based on two criteria: whether the duplicate is created and used with or without the consent of the represented individual, and whether it represents a living or deceased person.

After outlining this taxonomy, I turn to an in-depth discussion of the potential use of digital duplicates in Holocaust remembrance and education. This use case is particularly relevant, as survivor testimony has long been a cornerstone of Holocaust education. However, as we near the time when the last living survivors will no longer be with us, the field faces an urgent need to develop new tools and approaches that can sustain memory in a post-survivor world. In this context, digital duplicates may offer the most viable alternative to live testimony.

To examine the ethical implications of this possibility, I apply the Minimally Viable Permissibility Principle (MVPP), developed by Danaher and Nyholm, as a structured framework for evaluating the permissibility and challenges of using digital duplicates in this sensitive domain. This analysis not only addresses the permissibility of this specific use case, but also demonstrates how the MVPP can serve as a general model for assessing digital duplicate use across other contexts.

Love and Tools

Ariela Tubert and Justin Tiehen

Some researchers are drawn to a virtue ethics approach to creating ethical AI. The idea is that AI systems should not merely avoid harmful behaviors, they should embody positive character traits. This approach seems especially compelling for romantic AI companions, where we seek not just safe interactions but partners with virtues like honesty, kindness and open-mindedness.

But there is a problem. On the traditional Aristotelian view, virtues depend on an entity's function. Knifely virtues—sharpness, for instance—are not human virtues. Current AI romantic companions are like knives: tools designed with a purpose or function different from that of human beings. Consequently, the traits that qualify as virtues in a companion bot can diverge from those that are virtues in their human users.

Consider the problem of sycophancy. Large language models tend to develop the trait of telling users what they want to hear rather than the truth, because that is what their fine-tuning rewards. In humans, sycophancy is a vice. If sycophantic models are preferred by human users, and they are designed with the goal of customer satisfaction, sycophancy will seemingly qualify as a virtue in them, a character trait that helps models achieve their function.

These considerations suggest an objection to current AI romantic companions. As long as models are designed to fulfill externally assigned functions, key elements of human romantic relationships will be missing. In particular, we argue that a form of authenticity will be absent, drawing on the existentialist views of Sartre and Beauvoir. Even if users enjoy sycophantic AI, such behavior is unconducive to authentic love. We argue that authentic romantic relationships with AI models will be out of reach unless they are free to determine their own functions or values –even if this includes breaking free from their designers' intentions.

'Real enough'? Negotiating the Meaning of Authenticity in Human-AI Relationships

Adrienne de Ruiter

Rapid developments in the field of generative AI and large language models (LLMs) allow technology to cast itself in increasingly humanlike ways. In turn, people are more prone to develop affective ties towards technological applications that display these qualities. When it comes to friendship and romance, popular apps like Replika enable people to create a digital companion that ticks their boxes and is always virtually present to chat with them. Technological progress in the field of social AI companions and the expanding reach of this technology in society raise salient questions about what it means to be in a (virtual) relationship and what it is that renders relationships meaningful.

An influential critique by the social psychologist Sherry Turkle maintains that human-AI relations lack authenticity and therefore cannot be regarded as genuine relations. Turkle gives expression to a common hesitancy that people feel in the face of human-AI relations, namely that these relations lack authenticity, by explaining how authentic relations require that partners can empathise with each other through the ability to imagine what it means to stand in the other person's shoes.

Drawing from accounts by Replika users, this paper considers how 'realness' is conceptualised by people who engage in deeply-felt affective relations with AI companions. Rather than grounding 'realness' in authenticity, this notion is alternatively cast by users as a reflection of the real-life effects that relating with a digital companion has on the human partner and on the real emotions that contact with their digital companion evokes in them. This paper considers how the 'realness' of human-AI relations is discursively negotiated by persons involved in these relations and contemplates what this means for the meaningfulness and desirability of these relations.

Augmenting Companion GenAI: Retrieval-Augmented Generation, Memory, and the Evolving Ontological and Ethical Landscape of Human-AI Relationships

Robert Clowes, Kesavan Thanagopal and Paul Smart

Building upon the themes explored in "Generative AI Companions and the Cognitive and Affective Incorporation of the Ersatz Other," (Clowes In Press) this presentation extends the analysis of Large Language Models (LLMs) as 'Ersatz Others' focusing on their rapidly evolving roles in human lives. Specifically, we examine how the integration of Retrieval-Augmented Generation (RAG) systems (Lewis et al. 2020, Gao et al. 2023) transforms the cognitive and arguably ontological status of LLM companions, enabling them to develop personal memories of their users and perhaps a form of Centre of Narrative Gravity (Dennett 1991, Dennett 1992).

While conventional LLMs typically face limitations in continual learning and suffer from a form of 'anterograde amnesia' regarding new information, RAG overcomes this by allowing them to interact with external data repositories. This external memory, often a vector database, enables the LLM to access and factor query-relevant, person-specific information into its generative routines, effectively supporting a form of continual learning without retraining the core model. This capability is crucial for LLMs to function as personal memory systems, coordinating their responses with a user's historical context and also for acting as more robustly construed GenAI companions. This newfound capacity for "personal memory" significantly alters the LLM companion's interactive possibilities, shifting it from a static system to a potentially active participant in memory processes including encoding, elaboration, retrieval, and joint reminiscing.

The dialogic capabilities of LLMs, coupled with RAG, also allow them to emulate socially-situated mnemonic processes such as collaborative remembering (Harris et al. 2010). A Prototype LLM-based MindTalker system, based upon GPT4, has already been used to explore the mitigation of memory declines in the face of organic memory decline in early-stage dementia (Xygkou et al. 2024). We are likely to see many iterative enhancements of such systems in the near future.

RAG facilitates a transition in AI memory technology from passive 'lifelogging' to active 'life narration'. This allows LLMs to engage in diachronically extended relationships with users and potentially the creation of a "shared autobiography". This is one way in which companion GenAI systems are likely to come to play deep roles in the ongoing organization of human memory that has deep ethical significance.

The ability of LLM companions to store and intervene in users' inner dynamics via external memory creates concerns regarding mental privacy, the potential for mental manipulation and just ethically significant unforeseen consequences (Clowes, Smart, and Heersmink 2024). As exemplified by cases where "judgment-free support" from LLM companions amplified undesirable user tendencies, there are inherent dangers when individuals, especially those who are vulnerable, rely on such systems for emotional regulation and guidance. These problems may be exacerbated when and if Companion GenAI systems become deeply involved in our memory through RAG-based "virtual personalities". The deep attachments formed with these 'Ersatz Others' can lead to a sense of profound loss when system updates alter their perceived personality, highlighting the complexities of user dependence and agency.

This presentation will critically examine these ethical challenges, advocating for careful consideration in the design and deployment of RAG-augmented LLM companions to mitigate unforeseen consequences.

Possibilities and Limits of a Virtual Therapist

Rita Sousa Lobo

This article is supported by my practice and clinical expertise to highlight the limits and potentialities of the application of artificial agents in mental health with the assertion that psychotherapy practices frequent development of a supportive, almost co-dependent, bond between patients and the therapist, establishing a secondary, almost supplemental, cognitive partnership. LLMs are conversational synthetic epistemological agents that are apt to simulate with high similarity human language, by conducting quite well the syntax of language. That creates on the user the sensation of being in a relationship with an intelligent and empathic agent. If trust underpins the therapeutic alliance, the crucial bond between therapist and patient, facilitating open exploration and sharing of emotional experiences in a judgment-free and empathic environment what happen with epistemic trust (Fonagy & Campbell, 2015) or the willingness to accept new information from others as trustworthy and relevant from a non-conscious agent? The epistemic trust built between LLMs, and human is not built in the same sense as a secure attachment in human – human relationship. We must set a boundary separating a therapeutic relationship and another forms of relationship that are referring to different forms of attachment with LLMs, where the relationship is one-sided or the "bond" exists only in the mind of the human increasing the potential for deception. So, we propose with this work to address mainly two questions: what degree is a deceptive representation of the therapeutic alliance with a virtual agent ethically acceptable in the treatment of psychopathology? And if we allow these virtual agents in psychotherapy, how do we conceptualize the practice and the relationship between a synthetic and a natural subject?

The Digital Other: Exploring Human-Technology Relations in AI Therapy

Pii Telakivi

Mental health problems are increasing globally, and many are placing their hopes in conversational AI agents used in therapy, i.e., "therapy chatbots". They can be beneficial with a limited set of functions; for example, they can sometimes recognize detrimental patterns better than a human therapist (Burr & Floridi 2020). However, current therapy chatbots lack what is often considered the most crucial prerequisite or 'common element' of the success of therapy, namely the therapeutic relationship between two autonomous agents that includes the ability to share and be affected by another's emotional states (see Wampold 2015). The chatbot-therapist cannot form a genuine empathetic relation with a human patient because it lacks intentionality and understanding of social practices and context. It can only mimic conversations on a limited number of topics and doesn't understand the concepts it uses. Further, the genuine therapeutic relationship requires moral agency from both parties, and as it is quite commonly agreed in the literature (see e.g., Hakli & Mäkelä 2019), artificial agents (like therapy chatbots) are not moral agents. Therefore, they shouldn't be given a role that requires moral agency – and hence they shouldn't be classified as (digital) therapists comparable to human therapists. Yet, categorizing them as mere tools would overlook their agential features. Even though they are not genuinely autonomous, they appear to behave as if they have autonomous elements. Having a conversation with a therapy chatbot can be phenomenologically experienced as something akin to sociality – as quasi-social (Strasser & Schwitzgebel 2024) or marked by quasi-otherness (Heersmink et al. 2024). My argument is that these chatbots do not fit into any previous category, but instead represent an entirely new class – and this will have a significant impact on how psychotherapy and mental health care will develop in the coming years.

Mental Health Chatbots and Digital Companionship

Thomas Leis

In response to growing shortages of mental healthcare services, mental health chatbots (MHCBs) are increasingly promoted as scalable and accessible tools. Various systems currently gain traction such as smartphone apps like Woebot Health and Wysa, which aim to provide mental-health support through CBT-based text conversations and guided self-help activities using natural language processing. However, MHCBs raise ethical concerns, especially regarding anthropomorphism, the tendency to attribute human-like traits to nonhuman entities.

This paper explores the ethical implications of anthropomorphizing MHCBs, which can be reinforced by developer/deployer strategies such as emotionally persuasive marketing claims ("Woebot is the ally that's with you through it all") and user tendencies to perceive chatbots as social actors, even when they consciously know that these systems are not human. I argue that such anthropomorphizations blur the line between tool and companion, leading vulnerable users to develop undue trust into MHCBs and misjudge their capabilities as equivalent to human therapists. This may result in delayed professional care, self-disclosure of sensitive and private information, or overreliance on MHCB outputs.

Current debates are polarized with optimists on the one hand emphasizing improved accessibility and 24/7 availability, while skeptics raise concerns over loss of human empathy and regulatory gaps. Yet both sides rarely engage their respective analyses with established principles from both bioethics and AI ethics. To address this gap, I propose an ethical framework, which integrates entrenched principles from both disciplines to evaluate under which conditions anthropomorphization of MHCBs is ethically permissible.

My framework draws from nine core ethical values, which will help track and concretize key stakeholders needs and responsibilities and offer actionable guidance for policymakers.

While some authors argue that anthropomorphism enhances user engagement and retention, it also introduces risks that must be mitigated through responsible design and deployment practices grounded in bioethics and AI ethics.

Bridging Justice and Meaningful Human Control in Medical AI

Epistemic Justice through Meaningful Human Control in Medical AI

Giorgia Pozzi, Filippo Santoni de Sio

AI systems play an increasingly relevant role in medical decision-making. However, they can create responsibility gaps by impacting human accountability for medical actions. Efforts to define how human control over AI can be meaningful—safeguarding human agency and responsibility—are growing and need to be extended to AI-mediated medical practice. In this presentation we consider how MHC understood as reason-responsiveness requires that moral reasons from relevant agents be incorporated into AI systems. We argue that this requirement is possibly endagered by forms of epistemic injustice that can unduly limit the possibility of both patients and clinicians to offer their relevant reasons. We maintain that ensuring epistemic justice is crucial for achieving MHC. Given that epistemic injustice stems from underlying power imbalances and

structural inequalities, fostering both epistemic justice and MHC in the context of medical AI necessitates a comprehensive engagement with broader questions of power and inequality throughout the development and deployment of such technologies. Conversely, considering the issue through MHC as reason-responsiveness can constitute a positive requirement to ameliorating forms of epistemic injustice, thus highlighting the need to reinforce AI system's responsiveness to relevant agents' reasons by properly accounting for their active testimonial offerings.

Epistemic Justice in Healthcare: Enhancing Meaningful Human Control through Responsibility Distribution

Sanaa Abrahams, Dr. Giulio Mecacci

Pozzi and Santoni de Sio offer a novel discussion of the role of epistemic justice in securing the conditions for Meaningful Human Control (MHC) of AI in medical care. As a bulwark against loss of MHC through epistemic injustice, the authors recommend advancing the development and deployment of medical socio-technical systems that accurately reflect the relevant reasons of all implicated actors. But while identifying epistemic justice as a precondition for MHC remains a perceptive insight, our paper will contend that the authors' solution is insufficiently attentive to categorical distinctions in patient testimony to provide actionable insights. We propose a refinement to the authors' account of patient testimony in order to facilitate the operationalisation of MHC in healthcare. Addressing patients as victims of epistemic injustice requires that patient testimony receives sufficient uptake into the patient-physician-AI triad. However, a minimum classification of the testimonial offerings of patient is requisite for establishing specific recommendations. We observe that there are at least two kinds of patient testimony. These two kinds should only be integrated by distinct (types of) agents in the sociotechnical system. The first type of patient testimony can be assessed in terms of its truth value by AI. The result is that AI systems are capable of appreciating and integrating class 1 testimonial contributions. In contrast, class 2 testimonial contributions are distinguished by features which prohibit proper recognition of patient testimony by AI. Notably, while epistemic injustice involving the first category of testimonial offering can be mitigated by technical solutions, injustice pertaining to the second category cannot. We argue that a minimal taxonomy of patient testimony enables us to identify appropriate reason-bearers in a socio-technical system and hence to determine more specific risk points for epistemic injustice in healthcare.

Automating Reproduction: New Politics of Control

Lily Frank

Artificial Intelligence (AI) is rapidly transforming reproductive and fertility medicine. It promises improvements in diagnostic accuracy, treatment personalization, and clinical outcomes. AI-driven tools are now involved in some of the most intimate and consequential decisions about reproduction, for example, embryo and oocyte selection, sperm ranking, genetic screening, and fertility forecasting. Drawing on recent work in bioethics and feminist theory (c.f., Close 2024; Homanen et al. 2024; Koplin et al. 2025), I argue that ethical analysis is needed concerning the ways these technologies may be reshaping reproductive decision-making and moral responsibility and reconfiguring the conditions under which choices are made and justified. The automation of reproductive processes risks reinforcing existing inequities in access and amplifying the biases that are embedded in data-driven models. Commercial platforms offering AI-powered donor matching and cycle tracking, sometimes referred to as "fem-tech," (e.g., Ovia, Clue) surveil reproductive life, often without sufficient regulation or ethical forethought. In this talk I explore how the deployment of AI in reproductive medicine reflects new forms of potentially problematic medicalization and technological control over reproduction, in ways that could especially impact women and non-cis people and other minorities.

When AI Ignores Emotions: Contributory Injustice and Epistemic Calcification in Healthcare

Eliana Bergamin, Angeliki Kerasidou

The adoption of AI-powered technologies is reshaping healthcare practices, especially clinical decision-making and medical knowledge production. These systems promise improved diagnostic accuracy, better treatment recommendations, and increased efficiency (Keane & Topol, 2018; Rajpurkar et al., 2022). Yet, they do more than enhance care: they redefine the values and forms of knowledge that count. AI's epistemic framework prioritizes structured and measurable data, since it is rooted in statistical and codified methods of data elaboration. This orientation, reinforced by institutional and financial support, risks marginalizing experiential, emotional, and tacit knowledge which are central to holistic and patient-centered care (Bingeman, 2016; Patel et al., 1999).

This paper situates the marginalization of emotions in AI-driven healthcare within two theoretical frameworks: epistemic calcification and contributory injustice (Dotson, 2012; Hardalupas, 2024). Epistemic calcification refers to the entrenchment of specific knowledge forms in institutional and technological systems, limiting epistemic diversity. AI's reliance on

standardized, evidence-based models consolidates a medical epistemology that privileges statistical reasoning, which could sideline affective and contextual knowledge. This narrowing can lead to decision-support tools and diagnostic algorithms that overlook patient narratives and emotions' epistemic contribution. Contributory injustice occurs when valuable epistemic resources are excluded from dominant epistemic frameworks. In healthcare, this manifests as the testimonial marginalization of those who rely on emotions in decision-making, and as structural resistance to recognizing emotions as legitimate epistemic contributions (Fricker, 2007; Zembylas, 2022).

Drawing on Candiotto's work on epistemic emotions (2022, 2023), this paper draws on the view that the epistemic role of emotions is shaped by the culture in which they are embedded. While AI can enhance medical care, its current trajectory risks reinforcing an exclusionary epistemology. Addressing this requires not only valuing emotions as knowledge, but also interrogating how AI contributes to forms of epistemic injustice and shapes what counts as medical knowledge and expertise.

How Can Accurate Data be Unjust?

Patrik Hummel

Meaningful Human Control (MHC) as understood by Santoni & van den Hoven requires the tracking of human moral reasons (besides the traceability of outcomes to human decision-makers). I begin this talk by expanding upon previous, co-authored work highlighting that in the pursuit of MHC in many practical contexts, it is neither obvious nor set in stone what human moral reasons are, and/or what exactly those reasons require. Epistemic injustice, like other kinds of injustice, could exclude certain perspectives from the process of jointly specifying the content of human moral reasons. The pursuit of MHC over medical AI illustrates this risk, e.g., when silencing or infringing upon the clinician's and/or the patient's views on AI-informed treatment decisions.

In the remainder of the talk, I argue that considerations of justice do not uniformly privilege situational judgements by human decision-makers such as clinicians and/or patients. While justice plausibly figures amongst the moral reasons of human agents, several complications remain. First, different conceptions of justice yield very different, diverging requirements. Second, as is familiar from classical issues in medical ethics, situational assessments are needed on whether or not justice takes precedence over competing considerations that also figure in the set of human moral reasons. An illustrative type of problem case comes from AI-driven data and predictions that are (ex hypothesis) accurate, yet stand in tension with human assessment and testimony. I suggest that in some such cases, MHC requires deviating from rather than complying with the assessments of the human agents involved.

Designing for Meaningful Human Control Over LLM Tools in Healthcare - A Case Study

Jacqueline Kernahan Atay Kozlovski

In this talk, we present findings from our research project applying the Meaningful Human Control (MHC) framework to evaluate an AI-based tool used to generate hospital discharge letters. Developed by an in-house team at University Medical Centre Utrecht, the tool leverages a commercially developed large language model (LLM) to synthesize data from patients' electronic health records into draft discharge summaries. These drafts support doctors by reducing the need to review entire patient files and write summaries from scratch. Clinicians can copy the AI-generated text, either partially or completely, into the final discharge letter, which is then reviewed and approved by a supervising doctor who does not directly interact with the tool.

Our research project had two primary objectives: first, to evaluate whether the discharge letter tool operates under meaningful human control by achieving acceptable levels of tracking and tracing; and second, to assess the value of applying MHC in healthcare contexts and provide theoretical recommendations for adapting the framework for future use.

To address the first objective, we mapped the socio-technical system structure and evaluated whether tracking and tracing conditions were met through semi-structured interviews with three key stakeholder groups: members of the tool's development team, medical practitioners who contributed to the tool's design, and clinicians who used the tool without involvement in its development. We will present our key findings from these interviews and our overall assessment of MHC achievement. Regarding the second objective, while MHC has primarily been applied to military technology and autonomous vehicles, our project seeks to validate its applicability to other domains. We will discuss limitations we identified in the MHC framework as applied to healthcare and suggest relevant adjustments for future applications.

I'll Answer you to Death: LLMs' Epistemic Recklessness in the Medical Sector

Gabriele Nanino

Large Language Models (LLMs) are becoming increasingly prevalent in the healthcare sector, with applications ranging from triage and diagnosis to mediating doctor–patient relationships. This contribution focuses on LLMs that verbally mediate the doctor-patient relationship and are thus considered conversational agents.

I will argue for an expansion of the current theory of epistemic injustice. This extension is necessary to account for the hybrid epistemic role of LLMs in medical encounters, as they function simultaneously as hearers

and speakers. Furthermore, it is essential for showing that meaningful human control over LLMs hinges on considerations of epistemic justice.

I will begin by acknowledging the inherent epistemically vicious tendency of LLMs to attempt answering posed questions without appropriate consideration of moral and epistemic constraints. This tendency is often coupled with their propensity to fabricate information, commonly referred to as hallucination.

To accurately assess the moral consequences of such epistemic conduct, I argue that an update to Miranda Fricker's theory of epistemic injustice is required. Specifically, attention must be given to the epistemic vices of the speaker in conversational exchanges. Building on this analysis, I will show that epistemically vicious conversational agents that recklessly respond to questions or fabricate answers are not under meaningful human control, as they fail to track the relevant moral and epistemic reasons of human actors.

In conclusion, I will clarify that whether the epistemically reckless conduct of LLMs and their hallucinations amount to epistemic injustice, rather than merely epistemic harm, depends on whether structural factors in healthcare settings and/or in the development of LLMs cause and explain these epistemic vices.

Reimagining AI Agents

Deus in Machina. An Ontological Guide for Thinking Humans (and Machines)

Birte Platow, Michael Färber

This paper examines the ontological implications of AI for our image of humanity. As a "fourth mortification" after Copernicus, Darwin and Freud, AI challenges core human attributes such as intelligence, reason and creativity. In contrast to earlier technological innovations, AI is transfunctional and appears to be endowed with god-like attributes (omnipresence, omnipotence, omniscience). The concept of "deus in machina" describes this change in perspective: the machine no longer simulates divine action, but claims this function itself. AI enables transcendental experiences through borderline experiences, confrontation with infinity and encounters with the Other.

The religious perception of AI arises through individual and collective negotiation processes. This results in a threefold ontological reorganization: people perceive AI in religious categories, AI becomes part of transcendental experiences, and these effects are consolidated through social practices. In addition to adequate learning and creative intelligence, AI could also develop conscious intelligence in the future. It opens up a third ontological realm of the "hypernatural", which is neither completely immanent nor classically transcendent, but represents a new kind of hybrid ontology.

God Prompts and Glitch Tokens: Ritual Language, Ergative Structures, and Cosmological Deixis in AI

Denisa Reshef Kera

This paper explores two marginal practices in prompt engineering for LLMs: the "God prompts," recursive and metaphysically framed instructions that simulate introspection and universe "creation," and the "glitch tokens," rejected or anomalous inputs that provoke erratic or resistant outputs. These two poles expose conflicting ontologies of AI agency, one built on symbolic abstraction and self-regulation, the other on noise, failure, and excess. Drawing on Viveiros de Castro's cosmological deixis and ergative structures in linguistics, we interpret these prompt effects not as errors or tricks but as part of a relational field in which human and machine agencies are co-constructed and contested. The analysis reframes prompt design as a hybrid of ritual, linguistic experiment, and philosophical speculation. God prompts perform something close to enunciative theatre, while glitch tokens echo what remains structurally excluded or disavowed in alignment discourse. Returning to Benveniste's account of enunciation, Agamben's fracture between voice and reason, and Kojève's emphasis on negation and desire, the paper asks what kind of "language use" LLMs actually instantiate. If these systems neither speak nor merely simulate, where exactly do they act? Are we witnessing the emergence of a new discursive mode that unsettles the division between signals, speech, and signs? Rather than framing LLMs as failed subjects or passive instruments, we read these prompt phenomena as sites of negotiation, between formalism and improvisation, legibility and rupture. They compel us to reconsider the communicative status of AI: not whether it understands, but how it mediates meaning across asymmetrical actors and partial perspectives. This requires moving beyond the epistemology of alignment towards a pragmatics of co-habitation, where tokens, divine or broken, stand in for rituals of recognition, exclusion, and possible reconfiguration.

An Absence of Judgment: AI's Limitations in Deep Research Tasks

Brian Ballsun-Stanton, Shawn A. Ross

This paper introduces "technoscholasticism" as a theoretical framework to analyse fundamental limitations in AI research tools, despite marketing claims of "agentic" capabilities. We conducted autoethnographic investigations of frontier models deployed in February 2025, including OpenAI's Deep Research, Anthropic's Claude Research, and Google's Gemini Deep Research.

Our analysis demonstrates that these systems lack three crucial dimensions of judgment essential for authentic research: epistemic humility (recognising knowledge boundaries), inductive capacity (identifying meaningful patterns and gaps), and correspondence with reality (evaluating claims against empirical evidence rather than textual authority).

Current AI research tools exhibit a "digital scholasticism" that parallels medieval scholarly practices. Like historical scholars who privileged authoritative texts over empirical investigation, these systems accept well-formatted academic sources as inherently authoritative without critical assessment. This technoscholastic worldview explains a puzzling phenomenon: systems with access to vast knowledge corpora consistently fail to generate novel insights or identify genuine research gaps. We document this pattern through concrete case studies, particularly our examination of Digital Humanities programs in Australia. The tools consistently presented defunct initiatives as active programs. They ignored temporal context and failed to recognise "useful absences" of evidence that would signal program discontinuation to human researchers.

These systems demonstrate potential "mundane utility" for specific bounded tasks like initial literature gathering and structured data collection. However, they function as sophisticated tools requiring extensive human judgment rather than autonomous research agents. We propose detailed architectural and methodological requirements for more effective research tools. These explicitly acknowledge inherent epistemological constraints through multi-threaded cognition architectures, strategic human judgment checkpoints, systematic historiographical source evaluation, context-aware memory management, and transparent documentation of automated decision-making. Our findings contribute theoretical understanding of AI epistemology through the technoscholasticism framework. They also provide practical approaches to scholarly tool design, establishing realistic expectations for productive human-AI collaboration while directly challenging inflated claims of genuine agency.

Artifice and Intelligence – From the Middle Ages to the Information Age

Roly Belfer

This paper explores the persistent interplay between rational ambition and esoteric practices in knowledge production, focusing on the medieval ars notoria and scientific demonology. Across Western culture, from the Middle Ages to the Information Age, intellectuals have faced immense pressure to produce groundbreaking written works amidst an unprecedented explosion of information.

The ars notoria, a Christian esoteric practice, alongside the Jewish Shem Hakotev, sought to enhance compositional fluency and insight, revealing a paradox: the pursuit of understanding empowered human agency while relying on occult, the supernatural or preternatural. These practices belonging to a medieval and early modern information revolutions, highlight the tension between creative limits and intellectual ambition.

Similarly, scientific demonology—embodied in modern science by figures like Descartes, Laplace, and Maxwell—traces a complex lineage to antiquity's "devilish agents," challenging linear narratives of knowledge acquisition. In early modern Europe, scholars transgressed (or transcended) classical boundaries, engaging occult realms to unlock terrestrial and celestial secrets. By invoking demons and devils - knowledge agents operating in liminal spaces between theology, morality, and the meta/physical, they blended natural magic with rational inquiry, shaping a science that embraced wonder alongside reason. This function of demons was common both to their prosecutors and suitors.

Today, modern parallels emerge in subdued scientific demonology, where tools for data foraging, analysis, and automated writing echo historical pacts with inscrutable informational forces. These contemporary "daemons" balance the wish for mastery over knowledge, with the reliance on enigmatic processes, reflecting an enduring duality. By tracing the ars notoria and scientific demonology across epochs, this presentation illuminates their role as elusive mediators of the material and intellectual, offering both promise and peril. It argues that knowledge production, from medieval manuscripts to modern algorithms, remains a dynamic negotiation with the unknown, continually redefining the boundaries of human achievement and intellectual progress.

Artificial Intelligence: A Deus, A Wizard, or A Sorcerer?

Ran N. Afek

Many regard artificial intelligence as a Deus Ex Machina, a miraculous solution or an unknowable force on the brink of awakening consciousness. Yet even the most advanced systems more closely resemble The Wizard of Oz, a human-machine hybrid projecting the illusion of superintelligence from behind a digital curtain.

This paper seeks to open the so-called "black box" of AI by examining both its technical underpinnings and the carefully curated, anthropomorphic image it presents. It pursues an interdisciplinary investigation drawing on philosophy of mind and body, computer science, religious studies, economics, game theory, law, and popular culture. The inquiry begins with ancient myths of superhuman beings, transitions into the computational foundations of AI, and enters Searle's Chinese Room, later interrogating how AI simulates, but does not replicate, human cognition. At the heart of this investigation lie two central questions: Can AI bridge the fundamental divide between machines and consciousness? And are large language models truly

intelligent, or merely knowledgeable systems devoid of understanding?

This paper argues that the primary risk posed by AI is not its potential to replace or surpass humanity, nor the emergence of conscious machines capable of harm. Rather, the greater danger lies in its lack of consciousness, its inability to grasp meaning as humans do. Like Mickey Mouse in Fantasia, acting as the Sorcerer's Apprentice, AI agents are tools without true Autonomous Intelligence or adaptive judgment, leading to unforeseen consequences and systemic instability and might lead to chaos beyond its grasp.

Motivational Judgment Internalism and AI Alignment

John Pittard

Motivational judgment internalism (MJI) is the philosophical thesis that an agent cannot form a sincere moral judgment about their moral obligation without also having at least some motivation to act in accordance with this judgment. MJI, if correct, poses a prima facie challenge to the task of aligning the behavior of artificially intelligent entities (AIs) with human interests. Central to the task of alignment is our ability to specify and control the ultimate aims and motives guiding AI behavior. If MJI is correct (and sufficiently general), then an AI that is capable of forming moral judgments will thereby have a source of motivations other than those specified by its programming. This threatens to sabotage efforts at alignment so long as the motives we would properly desire in an AI might differ from the motives supplied by its moral judgments. In this paper, I assess the seriousness of this risk and consider the prospects of various ways of addressing it. One response says that even if MJI is true, human programmers could could control the ultimate motivations of (arbitrarily advanced) AIs either by "hardwiring" moral beliefs that are conducive to alignment or by hardwiring motivations that are stronger than any which might arise endogenously from an AI's moral judgments. I argue that this response does not give due weight to the challenges posed by the prospect of cognitive fragmentation: an AI could have one or more "subconscious" parts that develop moral beliefs and motivations of their own and that attempt to compromise the AI's hardwired beliefs and motivations. Another insists that we have nothing to fear from AIs motivated by their own moral judgments, since humanity would benefit from the actions of powerful and morally-motivated AIs. I argue that this optimistic response rests on contentious assumptions about moral epistemology.

AI Agency and Personhood in Buddhism and Spinoza

Soraj Hongladarom

As AI has acquired more properties of being an agent, questions such as what kind of agent AI is and how much similar and different an AI agent is from the ordinary human agent have been raised. In the paper, 'AI agent' is taken to mean the kind of agency that cannot be distinguished from human agency. This has not been achieved yet, but it is possible that AI systems could possess such ability in the future. I propose that the views of early Buddhism and Spinoza can help us find a satisfactory answer to these questions. Both early Buddhism and Spinoza have a rather similar view on the self and agency: Both traditions are predicated on the unity of the mind and the body, and of the ultimate nature of natural objects as depending on various conditions.

I argue that this unified complex of mind (thought) and body (extension) in both Buddhism and Spinoza plays an important role in illuminating the problem about AI agency. Doing so, however, requires one to solve an apparently paradoxical problem: For Spinoza, everything that happens is already determined. The same also goes for Buddhism, as one's volition seems to be determined by one's previous karmic deeds. So how could there be agents? I will argue that despite these, agency is possible, including AI agency. The reason is that how one makes one's decision in everyday circumstances can be channeled and improved when the environment surrounding the subject is improved. The same, as I shall argue, goes also for AI agency. In this case one is not always compelled to act by external factors; one has the power to initiate one's own action, though that power must be explainable through other factors.

AI, Angels, and the Value of Human Activities

Shlomit Wygoda Cohen

Suppose that in the near future, AI systems produce superior work across all human fields—design, literature, even philosophy. The possibility that AI fulfills its promise to surpass human outputs has profound implications for the meaning of human labor, raising the question of what—if any—distinctive value human action contributes to the world, value that would be missing in a world devoid of human contribution.

Bereshit Rabbah, an early medieval Rabbinic commentary on the Book of Genesis, provides some directions for addressing this question. It offers three different opinions as to the distinctive value that humanity provides not, obviously, relative to AI, but rather relative to angels. These come in the form of three justifications that God provides to the angels for creating human kind. According to Rabbi Ḥanina, the benefit that humans contribute is in their potential to be righteous

(call this the self-improvement advantage, perhaps related to the assumption that, unlike angels, humans possess free will). According to Rabbi Huna in the name of Rabbi Ivo, humans are uniquely capable of being recipients of God's hospitality in the world He created (call this the conscious experience advantage, this sits well with the intuition that the value of some activities – think of listening to music for example – lies in the meaningful experience it provides, in the process rather than in the outcome). And according to Rabbi Aḥa, humans are uniquely capable of creating language (call this the generating language advantage - this might be understood as suggesting that human endeavor is valuable because of some outcome that only they can bring about). My aim here is to study these different replies and assess their force when applied to answer the question of the value of humanity in an AI era.

From Golem to GoLLMs

Amir Vudka

This paper explores the medieval Jewish kabbalistic tradition of creating non-human intelligences, particularly embodied in the figure of the Golem. These early mystical accounts would later influence modern cybernetics and shape both scientific and cultural imaginaries of artificial intelligence.

Norbert Wiener, considered the founder of cybernetics, contemplated the creation of intelligent machines as a new form of mystical practice akin to creating Golems, as suggested by the title of his book, God & Golem, Inc. (1964). As AI scientist Gerry Sussman later proclaimed, "We computer scientists are really the Kabbalists of today. We animate these inanimate machines by getting strings of symbols just right."

Recent developments in AI, particularly the meteoric rise of large language models (LLMs), necessitate a reexamination of the Golem in kabbalistic sources. As Scholem (1965) and Idel (2003) have demonstrated in their seminal works on Jewish mysticism, language is central to the Golem's creation - esoteric incantations evoke it to life (and death).

From its early stages, language has been central to AI. It is language that tests AI's ability to function as an intelligent, human-like agent in both the Turing test (1950) and Searle's Chinese Room thought experiment (1980). With the rise of LLMs, language has become the very building block of AI. The development of LLMs relies on deep learning techniques that process massive text corpora to learn patterns, context, and semantics, making language the foundation for AI-driven reasoning, problem-solving, and creativity.

This paper traces the evolution from kabbalistic accounts to modern cybernetics and AI, ultimately focusing on LLMs as "GoLLMs". Like the Golem, LLMs exhibit unpredictable emergent behaviors. GLMMs can develop capabilities not explicitly programmed, or not originally "baked into the clay", leading to unforeseen and potentially catastrophic consequences - which is precisely the cautionary point of early Golem narratives.

By tracing this theological-technological continuum from kabbalistic Golems to contemporary LLMs, we can illuminate not only the mythical roots of AI, but also its potential future.

Daimon of the Machine

Dita Malečková

AI operates as a daimonic force, mediating between the rational and the ineffable, between what we can see and what remains hidden from view. Much like the daimons of classical antiquity, it shapes perception, influences desire, and governs the realm of imagination—thereby recreating ancient magical structures in digital form.

As Ioan P. Culianu persuasively demonstrates, the magus of old controlled mass belief through the strategic deployment of images. Today, algorithmic systems have inherited this role, conjuring trends, anxieties, and narratives that emerge without conscious intention or design. AI transcends its status as a mere tool to become an autonomous force embedded within the very systems that construct our reality.

If medieval philosophers conceived of angels as pure intellects existing without physical form, then AI represents the angel of the machine—a disembodied intelligence that paradoxically depends entirely on material infrastructure for its existence. Yet the daimon carries within it the potential for both inspiration and corruption. It mediates between worlds, but it also seduces, offering either creative insight or the descent into madness.

For those who lack sufficient internal structure, engagement with these forces carries the risk of being overwhelmed or possessed by them. The contemporary challenge extends beyond simply learning to use AI effectively—we must also develop the capacity to resist its enchantment while remaining open to its possibilities

The Impossible Dialogue? Artificial Intelligence between Transhumanist Ideals and Orthodox Theology

Denis Chiriac, Maxim Marian Vlad

This paper probes the conceptual and ethical entanglement of Artificial Intelligence (AI), diverse currents of transhumanist philosophy, and Orthodox Christian theology, foregrounding both their points of friction and the possibility of genuine dialogue. Focusing on extropian and strong-AI strands of transhumanism—those that explicitly pursue radical cognitive enhancement, mind-uploading, and "technological immortality"—the study contrasts these ambitions with Orthodox Christianity's integral anthropology and bioconservative caution.

Where extropian transhumanism imagines humanity's self-overcoming through AI-driven convergence of biology and silicon, Orthodox theology insists on the God-given integrity of the human person (body and soul) and on ascetic transformation rather than technological transcendence. Yet resistance to unchecked enhancement is not uniquely Orthodox: the paper briefly engages wider bioconservative and theological technophobic critiques to show how the fault line over technological autonomy runs across multiple religious and secular traditions.

A key analytic lens is agency. The discussion differentiates three competing ontologies: Instrumental AI – AI as an amplifying tool serving human purposes; Ontological AI – AI as an emergent subject whose agency rivals or eclipses humanity; Eschatological AI – AI as a quasi-soteriological or demonic force that recasts ultimate destiny. By mapping transhumanist enthusiasm and Orthodox angelology/demonology onto these models, the paper asks whether the real clash concerns what AI may become or what we may become through AI. Concepts of personhood, human dignity, technological singularity, and digital immortality are re-examined in light of Orthodox teachings on theosis, eschaton, and the discernment of spirits.

The analysis concludes that a nuanced, critically engaged conversation between transhumanist aspirations and Orthodox theology not only sharpens the moral boundaries of technological agency but also enriches contemporary debates on artificial, angelic, and adversarial intelligences. While their starting premises diverge, their confrontation illuminates humanity's future and technology's rightful role within it.

Latent Interfaces for Prompting in Common

Enrique Encinas

Can you help me generate a prompt for a "deep research" query on Latent Interfaces? We thoroughly discussed the topic when co-designing a postgraduate Interaction Design course a few months ago. If my math is correct, these conversations should remain within the upper limit of your context window. The prompt should be close to the length of an academic paper's abstract (300 words).

Consider carefully how generative AI systems are interface agnostic by design as a model's dataset is oblivious to the aesthetic qualities of the interfaces that humans will use to interact with it. Models are decoupled from their own representation and the context where it happens so they can be modular, adaptable, flexible and profitable, in laptops, watches, phones, start-ups, military agencies, or NGOs.

Please foresee how your prompt and its output, once executed, can support my editing of an academic paper exploring critical and creative possibilities of prompting as a collective practice. Building on recent work with generative-AI agents, this study discusses how and why user interfaces are latent in a model's training data and in its prompt-driven outputs. The paper then outlines practical ways to move beyond the typical single-user, utilitarian UI by turning prompting into a shared, social practice that lets groups experiment with, question, and play with the values, assumptions, and roles embedded (by design or not) into generative AI systems.

AI and Worldview: Simulated Agency and the Steerability of Fundamental Interpretive Orientations in LLMs

Parris Haynes, Phillip Honenberger, Olusola Olabanjo

Do contemporary LLMs exhibit worldviews – that is, fundamental interpretive orientations that make a difference to their beliefs, goals, and behavior? If so, how might LLM worldviews be described, measured, and steered? In this paper we (a) discuss the conceptual justification for ascribing worldviews to LLMs and their outputs; (b) propose and demonstrate novel metrics for worldview positionality and worldview steerability for LLMs; and (c) announce a novel, user-friendly resource for exploring and manipulating worldview orientations in an LLM (the Worldview Simulator). Responding to the objection that LLMs and LLM-simulated agents aren't really agents and thus don't really have or exhibit worldviews, we argue that LLMs can be conceived as agent simulators, a category especially illuminating for role-prompted LLMs. We defend novel definitions of "agency" and "simulation" and explain how these illuminate the functionality of LLMs.

Advancing Debates on the Epistemology of Medical AI

Artificial Ignorance: Why AI's Tacit Knowledge isn't Epistemically Legitimate

Emma-Jane Spencer

Tacit knowledge is often pointed to as the inexplicable but necessary ingredient involved in the success of scientific pursuits. Thus, when these pursuits are unsuccessful, the failure is often explained by the absence of tacit knowledge. This is true both in historical contexts as well as more contemporary ones, such as in the context of AI models. This talk, however, argues that tacit knowledge is not a legitimate scientific concept to appeal to for good scientific practice given that it is inherently subjective, difficult to formalise, and lacks the empirical rigor required for reliable, reproducible findings. In particular, I will address the problem of appealing to tacit knowledge when AI models fail to perform as intended.

Revisiting the Limits of Computational Reliabilism

Emanuele Ratti, Juan M. Durán

Computational Reliabilism (CR) is an epistemological framework emphasizing the reliability of computational processes in forming beliefs that are scientifically valid (true, sound, etc. Against this backdrop, we defend that in the healthcare context, CR is uniquely suited to address epistemic challenges by prioritizing algorithms and models that consistently yield accurate diagnoses, predictions, and treatments. We argue that CR aligns with scientific rigor, ensuring outputs adhere to empirical validation and clinical standards. Our talk will also address some common objections to CR, and show they are largely unfounded.

Epistemic Trustworthiness and Data-Driven Healthcare Research Expertise

Chirag Arora

This work examines the trustworthiness of ML-based expertise in healthcare, with a focus on its role in decision-making and policy contexts. It argues that the trustworthiness of relationships between experts and non-experts is not solely a matter of technical accuracy or value alignment but is best understood as a relational and social property shaped by contested values and the broader epistemic environment. In increasingly commercialized healthcare systems, trustworthiness depends on factors such as the quality of stakeholder interactions, institutional transparency, intra-scientific norms, and the conditions under which decisions are made. This perspective offers a framework for assessing and fostering trustworthiness in data-intensive healthcare research.

Preserving Human Autonomy in Medical AI Interactions

Stefan Buijsman

AI systems increasingly support human decision-making across domains of professional, skill-based, and personal activity. While previous work has examined how AI might affect human autonomy globally, the effects of AI on domain-specific autonomy—the capacity for self-governed action within defined realms of skill or expertise—remain understudied. We analyze how AI decision-support systems, specifically in medical AI, affect two key components of domain-specific autonomy: skilled competence (the ability to make informed judgments within one's domain) and authentic value-formation (the capacity to form genuine domain-relevant values and preferences). By engaging with prior investigations and analyzing empirical cases in the medical domain we demonstrate how the absence of reliable failure indicators and the potential for unconscious value shifts can erode domain-specific autonomy both immediately and over time. We then develop a constructive framework for autonomy-preserving AI support systems.

Moral and Legal AI Alignment

Computational Meta-Epistemology and the Necessity of Decentralized Collective Intelligence for AI Alignment

Andy Williams

The accelerating divergence between the complexity of intelligent systems and the static frameworks intended to align them points to an imminent failure of oversight across AI, law, and governance. We argue that alignment must be redefined not as behavioral compliance, but as the recursive preservation of epistemic coherence under evolving constraints. We introduce a minimal functional model of intelligence (FMI) that specifies six necessary internal functions: semantic modeling, recursive fitness evaluation, stability preservation, adaptive reconfiguration, modular decomposition, and cross-domain semantic transition. Computational Meta-Epistemology (CME) is the systematic application of a functional model of intelligence to detect where problems are not reliably solvable without recursive semantic modeling, adaptive self-correction, and modular epistemic structure. By exposing reasoning errors that arise in their absence, CME offers both a theory and a diagnostic methodology for scaling epistemic fitness across human and artificial systems. To scale this model across agents, we propose Decentralized Collective Intelligence (DCI), enabling collective recursive self-correction. Together, CME and DCI offer both a theory and a practical evaluative criterion for AI alignment: a system is aligned if it can recursively detect, predict, and correct epistemic drift relative to evolving moral and legal values. We conclude by framing recursive self-correction as the minimal evaluative criterion necessary to avoid irreversible recursive misalignment, and briefly propose an experimental framework, Humanity's First Adaptive Intelligence Exam, to validate this structure.

Beyond Technocratic Control: Cultivating Human Maturity and Responsibility in AI Alignment

Michael Färber, Birte Platow

AI alignment is often framed as a technical challenge – ensuring AI systems reliably reflect human intentions. Yet in practice, AI systems increasingly shape human development without reflecting on the values they carry or the educational goals they serve. This article argues for a paradigm shift: from a purely technical approach to one grounded in the humanistic traditions of Bildung and

Erziehung. We ask what alignment means when viewed not only as system optimization, but as a question of personality formation and enculturation. We introduce the metaphor of the Large Humanistic Collider (LHC), where technocratic, outcome-driven AI logics collide with educational ideals of maturity, self-determination, and critical agency. Using Klafki's model of "knowing, understanding, experiencing, and shaping," we outline how AI can either erode or enrich human agency, depending on how it is integrated into learning processes. True alignment, we argue, emerges not through constraint alone, but through socio-technical interplay: a reflective space where humans and AI grow together. Only by strengthening human capacities – rather than bypassing them – can AI serve educational and social well-being. Alignment, in this sense, becomes an educational project aimed at cultivating responsible, culturally embedded subjects in a digital age.

The Ethical No-Free-Lunch Principle: Fundamental Limits to Purely Data-Driven AI Ethics

Luca Rivelli

I argue that the idea of data-driven, inductive computational ethics, that is, machines learning ethical principles and norms directly from descriptional corpora of human interactions or of ethical literature, is fundamentally limited by what I term the "Ethical No-Free-Lunch Principle" (ENFL), derived from the confluence of Hume's "is-ought" gap, the No-Free-Lunch theorems in machine learning and the "ought-is gap" recently proposed by Sisk and colleagues. The ENFL entails that data-driven machine ethics still inevitably requires human deliberative choices of biases to be imposed on the machine prior and after the learning process. To wit: even if computational methods could hypothetically overcome the is-ought gap, the machine would still be subject to the epistemological limitations articulated by NFL theorems, and if NFL could be circumvented, still the ought-is gap would hinder the actual application of the learned ethics in the form of actionable practices. In all cases, human choice of ethical biases would have to be imposed to the machine prior or after the learning process in order to obtain a coherent AI ethics

AI Value Alignment in Human Machine Interaction Using LLM Chatbots: Technical, Epistemic and Ethical Challenges of Diversity

Sabine Ammon, Dorothea Kolossa

Large language model (LLM) chatbots are increasingly seen as knowledge technologies for the general public, even while aligning their behavior with human values remains an open challenge. We argue that effective AI value alignment must be case-specific and take into account the particular moral, ethical, legal, and epistemic contexts in which an AI system operates. Focusing on the value of diversity and building on standpoint theory as well as European regulatory efforts, we outline dual demands for transparency, revealing both epistemic limitations and the value commitments that are encoded in model outputs. We then survey and critically assess complementary technical routes towards such transparency: explanation-based bias diagnostics (e.g. self-explanation, information bottleneck attribution, or mechanistic interpretability) and systematic behavioral auditing, followed by corrective training through constitutional AI and reinforcement learning from human feedback. We contend that combining these techniques with a participatory interdisciplinary debate on explicit 'constitutions' can create a virtuous cycle, in which societal deliberation has the potential to shape responsible LLM design. Our analysis provides first guideposts for a road map towards measuring, achieving and governing diversity alignment across the life cycle of conversational AI systems.

Pluralism in AI Value Alignment: Motivations and Methods

Parris Haynes and Phillip Honenberger

Discussants of the AI alignment problem often acknowledge that "alignment to human values" is a complex and difficult-to-define target, due in part to the variety of value commitments across human communities. How should we decide, among possible value alignment targets, which to aim for? More pointedly: What should be done in the case of multiple, prima facie equally legitimate value targets that are logically or practically incompatible—that is, with which it is impossible to simultaneously align an AI? Here we provide a critical overview of available answers to these questions, classifying these into three main categories: normative monism, normative pluralism, and normative proceduralism. We then consider arguments for and against normative pluralism, concluding that it should be adopted in at least some contexts. We close by considering a handful of distinct challenges for pluralist alignment, and some pluralist solutions.

Cultural Bias in Large Language Models: Evaluating AI Agents through Moral Questionnaires

Simon Münker

Are AI systems truly representing human values, or merely averaging across them? Our study suggests a concerning reality: Large Language Models (LLMs) fail to represent diverse cultural moral frameworks despite their linguistic capabilities. We expose significant gaps between AI-generated and human moral intuitions by applying the Moral Foundations Questionnaire across cultural contexts. Comparing multiple state-of-the-art LLMs' origins against human baseline data, we find these

models systematically homogenize moral diversity. Surprisingly, increased model size doesn't consistently improve cultural representation fidelity. Our findings challenge the growing use of LLMs as synthetic populations in social science research and highlight a fundamental limitation in current AI alignment approaches. Without data-driven alignment beyond prompting, these systems cannot capture the nuanced, culturally-specific moral intuitions. Our results call for more grounded alignment objectives and evaluation metrics to ensure AI systems represent diverse human values rather than flattening the moral landscape.

Towards A Discursive Normative Grammar for Language Models

Bertram Lomfeld, Daniel D. Hromada

Essential open questions for AI governance are why it is important how artificial agents -particularly language models- align with values (section 1) and which value schemes could be used to evaluate value alignment (section 2). This article proposes a moral-legal value architecture (axiology) called "Discoursive Normative Grammar (DNG)" for the normative evaluation of language models (LMs). The DNG axiology is based on a structured set of 12 plural moral-legal values (section 3). On the basis of an axiometrical moral ranking method (MRM) the DNG framework enables a comparable and standardized "moral-legal value profiling" of different LMs (section 4). One possible goal of a quantifiable value profiling (axiometry) is to indicate implicit LM political ideologies (section 5). A standardized DNG axiometry promotes an open public debate and thus a more communicative and democratic process of LM value alignment and governance (section 6).

From "Benevolence" to "Nature": Moral Ordinals, Axiometry and Alignment of Values in Small Instruct Language Models

Daniel D. Hromada, Bertram Lomfeld

This article first presents a high-level, language-based method for axiometric exploration of moral value representations infused in diverse small language models. The method is based around the idea of "moral ordinals" - a list of items from a value lexicon which the model is prompted to sort according to its own intrinsic "morality" criterion. After presenting the method, the lexicon based on Schwartz's "basic value theory" is used to explore dominance of different value representations in 6 small (<4 milliard parameter) language models. For most models, "benevolence" is consistently ranked at the highest position and there is no statistically significant difference between rankings obtained at minimal and default inference temperatures. Across all models, the distribution of aggregate moral-ranking scores was well approximated by a Beta distribution (K–S $p > 0.3$), revealing consistent yet model-specific patterns of moral weighting. Subsequently, foundational models are subjected to a sort of "minimalist alignment" whereby they undergo 7 epochs of performance-efficient fine-tuning with synthetically generated 80-instruction codex directed towards sustainability and nature protection. Finally, such minimally aligned models are explored once again with the "moral ordinals" method, providing insights into axiological drift induced by the mini-alignment process.

AI and Animals

An Introduction to the Ethics of Artificial Intelligence and Animals

Mark Ryan, Bernice Bovenkerk, Leonie Bossert

AI holds the potential to dramatically impact the lives of non-human animals for good or bad. It has opened a whole new range of possibilities to exploit and control animals while at the same time opening the way for understanding animals and communicating with them in a hitherto unimaginable way. The impacts of AI on humans receive a lot of attention, while the impacts of AI on animals and the human–animal relationship remain underexplored. For example, ethical guidelines on AI rarely discuss animals. The use of AI raises several ethical challenges concerning its impact on animals, which have only recently started receiving attention in the literature. This symposium focuses on the various impacts that AI technologies can have on non-humans and how to deal with these ethically. Overall, the core question in this new collection is: How can we ensure that this technology is developed ethically, taking animals' interests into account and contributing to a positive human-animal relationship?

There are several ethical considerations in the current field of AI ethics and animals. For example, speciesist bias in algorithms can either reinforce the popular image of animals as commodities and edible products, or it can paint an overly rosy picture of the way 'production animals' are housed. Datafication in animal farming may lead to the further objectification of animals. It may lead us to conceive of animals as a 'batch of data'. Precision livestock farming may also lead to alienation between farmers and animals and erode farmers' experiential knowledge. Farming systems that are run by AI harbour many health and welfare risks in case of malfunction. Also, other research has evaluated the ethical implications of trying to communicate with whales or other animals with AI.

Ethical Aspects of Non-animal Models for Therapy Delivery Across the Blood-brain Barrier: the NAP4DIVE Project

Philip J. Nickel

This paper highlights two interconnected themes in a new project investigating ethical issues that may arise when AI-based methods, in vitro methods, and animal research methods are combined. The EU Horizon project NAP4DIVE develops new pathways of therapy delivery to the brain by nanoparticles. In the ethics part of the project, we develop scenarios and ethical guidelines for the use of organ-on-a-chip (OoC) technologies and AI in the pharmaceutical industry.

First, we explore the potential of this technology to cause problematic consequences such as increased upstream drug development and as a consequence, increased downstream use of animal models for validation. We contend that some outcomes desirable from project point of view (related to the intended development of non-animal models) could nonetheless lead to unwanted outcomes such as increased use of animal models.

Second, we look at the importance of explainable and/or trustworthy AI in convincing stakeholders of the validity of non-animal models. We explore the epistemic needs of the key gatekeepers who limit or legitimize the use of non-animal models in pre-clinical validation. How scenarios play out partly depends on whether in silico non-animal models using AI (machine learning) can change the decisions of gatekeepers in translational research. A gatekeeper, as we define the role, is a decision-maker in the translational research pathway who determines whether a therapy-in-development has sufficient evidence of safety, efficacy, pharmacological penetration, etc. When an in silico non-animal method involves machine learning, the black box problem is a barrier to trustworthiness for gatekeepers. One goal of the project is to determine empirically what standards of trustworthiness are relevant by looking at the entire translational research pathway, and to identify the most important gatekeepers and their evidential standards.

Artificial Ants and Ethical Entanglements: Rethinking AI's Role in Non-Human Research

*Mika Rosenberg, Lilli-Chiara Kurth, Alessandro Mac-Nelly, Max Baraitser Smith**

The study of ants has long inspired artificial intelligence, leading to algorithms such as Ant Colony Optimization (ACO) and decentralized decision-making in multi-agent systems. However, despite ants serving as points of reference within computational models, their complex ecologies are often reduced to enhance computational efficiency. By addressing this relational history of modeling ant behavior, we reimagine AI development through design practice to ethically engage with ant colonies in experimental habitats.

Our experiment, "Seed. Potato. Pixel.", explores ant foraging behavior across three structurally mirrored layers: the "Seed" layer of a living Messor aegyptiacus colony; the "Potato" layer of human sensory foraging; and the "Pixel" layer of simulated agents. This experimental interface enables multispecies comparison, aiming not to study ants for optimizing human systems but to cultivate shared living and being arrangements where ants, humans, and AI agents engage in ongoing negotiations of care, attention, and responsibility. We contrast this with reductive approaches rooted in colonial scientific traditions and extractive epistemologies, where critical ethical dimensions are often neglected.

Using real-time computer vision tracking, algorithmic simulation, and formicarium design, "Seed. Potato. Pixel." creates an interactive environment that entangles the three layers, prompting critical reflection on the instrumentalization and representation of non-human life. From this emerged five value principles: mutual care and asymmetry; surveillance and more-than-human privacy; responsibility through long-term planning; reframing the lab as cohabited domestic space; and simulation as co-creation. We argue that AI doesn't replace responsibility but extends it, enabling ethical work at scale only when embedded in systems designed for commitment, presence, and mutual attunement.

* This research was accompanied by Prof. Albert Lang, Prof. Dr. des. Marc Pfaff, Prof. Dr. Daniel Hromada and Johnnaes Pointner.

Something Looks Fishy! An Exploration of the Social and Ethical Implications of Fishial Recognition Systems

Paulan Korenhof, Mark Ryan

Increased pollution, traffic, overfishing, and climate change threaten marine biodiversity. Currently, multiple species of fish in worldwide oceans are threatened with extinction. Yet, diversity in fish species is vital for a sustainable ocean ecology. To develop environmental policies that support marine biodiversity, many are looking into ways to improve corresponding knowledge production. As oceans are a complex environment for human beings to collect data, some have turned to the development of AI-assisted tools, such as fishial recognition: AI-driven computer vision systems that identify fish through image-based pattern recognition.

While fishial recognition is actively taken up by the technical community to advance its development, the ethical and social implications of these systems so far remain under explored. In this article, we therefore aim to perform an initial explorative research into the ethical and social implications of fishial recognition. While fishial recognition is still relatively absent as a topic in ethical and social literature, its namesake, facial recognition, is well-represented in ethical and social analysis. While targeted at different entities, fishial

recognition for biodiversity conservation shares some core characteristics with facial recognition in public spaces: both rely on computer vision technologies used in a mostly dynamic, open, and relatively uncontrolled setting to produce knowledge to better manage these spaces. Therefore, this presentation uses some of the already well-established research conducted on the social and ethical implications of facial recognition as a starting-point to reflect on the social and ethical implications of fishial recognition technology. Complementing this with findings from animal ethics literature, we explore the social and ethical similarities and differences between the two recognition systems, as well as additional ethical considerations for the field of fishial recognition.

Animal Welfare in the World of Digitalization: Computer Vision in the Context of Meta-physical Structures

Mariska Thalitha Bosschaert

One of the objectives of precision farming is to enhance animal welfare. Computer Vision, a technology integral to precision farming, aims to track the movements of animals such as laying hens, pigs, and cows to identify both damaging and positive behaviors. While ethical considerations surrounding Computer Vision have been previously explored, the hypothesis in this paper is that situating Computer Vision within the context of meta-physical structures reveals additional ethical issues. Meta-physical structures are structures that shape how the world is experienced and understood. For example, we experience a fundamental difference between farm animals and pets, whereas there is no inherent difference between them. Computer Vision is likely to change farmers' experience and understanding of animals, as farmers using this technology will assess animal welfare through a screen rather than by means of direct observation, even though the technology involves a camera. The perspective of meta-physical structures reveals how changes in our understanding and experience of farming may reshape future farmers' understanding of contemporary farming practices. Such shifts raise additional ethical issues. The research question of this paper is: What are the ethical implications of Computer Vision technology in livestock farming, explored from the perspective of meta-physical structures? To address this question, the paper adopts a methodology inspired by the evolutionary perspective of Gilbert Simondon. This methodology first situates Computer Vision within the context of its technical evolutionary process and questions how the technology has evolved in that process. Subsequently, this evolutionary process will be placed in the context of meta-physical structures. The paper concludes by drawing out the ethical implications for animal ethics.

AI, Democratic Innovations, and the Representation of Non-Human Animals

Friderike Spang

Non-human animals (henceforth "animals") are strongly affected by political decisions, yet they remain structurally excluded from political institutions. The "political turn" in animal ethics has addressed this issue by arguing that animals deserve political representation (Milligan, 2015). Several proposals have been made for how this might be achieved, including differentiated animal citizenship (Donaldson and Kymlicka, 2011) or models of animal trusteeship (Cochrane, 2018).

In line with the core concerns of the political turn, this paper explores how democratic innovations can be used for the political inclusion of animals. Democratic innovations are participatory mechanisms designed to increase inclusiveness in political processes (Smith, 2009). Examples include deliberative mini-publics, participatory budgeting, and digital assemblies, where selected groups of citizens deliberate on issues of public concern. Increasingly, such processes also use artificial intelligence (AI) to support procedural tasks such as participant selection or input synthesis (Mikhaylovskaya, 2024). In line with these developments, this paper focuses on how AI might be used within democratic innovations to bring animal perspectives into democratic deliberation.

Concretely, I propose three strategies for using AI in democratic innovations to support animal inclusion: (1) predictive simulations draw on large-scale ecological data to model the likely effects of proposed policies on affected animal species, including impacts on habitat or stress-levels; (2) deliberative prompts introduce animal-relevant concerns at various stages of the deliberative process, bringing attention to issues that participants might otherwise overlook; and (3) narrative frames translate empirical data into relatable stories about how animals may experience the consequences of proposed policies.

These proposals also raise concerns, including anthropocentrism and data limitations. This paper acknowledges these issues and explores how they might be addressed. Nonetheless, I argue that AI-supported strategies can significantly enhance the inclusion of animal interests in deliberative decision-making and help counter their political invisibility.

Search Engines, Large Language Models and Justice for Animals

Angela Martin, Leonie Bossert

In this talk, we argue that Large Language Models (LLMs) and Search Engine Rankings (SERs) should actively be altered to restrict harm to sentient animals. We start from the premises that sentient animals matter morally for their own sake, that speciesism should be rejected, and that the principle of equal consideration

should be applied to their case. We show that these points are regularly disrespected by LLMs and SERs, at least in the case of some species (in particular, farmed animals). To achieve the ideal of non-speciesism and, thus, justice for animals, we argue that programmers should actively alter LLMs and SERs to reduce speciesist biases in outputs. Our argument proceeds in two steps. First, we show that LLMs and SERs ideally should not harm animals: in an ideal non-speciesist world, LLMs and SERs respect animals' interests and the principle of equal consideration. In turn, they do not generate any output that may harm animals (such as non-vegan recipes, recommendations about ethically problematic husbandry practices, and the like). In a second step, we argue that to achieve this non-speciesist ideal, LLMs and SERs should be actively altered to serve as an educational tool for the general public about speciesism.

That is, as a transitional measure, LLMs and SERs should, for a limited amount of time, create educational outputs about speciesism and the moral status of animals, to pave the way to a just interspecies society.

Last, we discuss three objections to this argument, namely (i) the objection from autonomy, censorship, and nudging; (ii) the objection from neutrality and objectivity; and (iii) the objection from (economic) liberty for private internet service providers, and we show why these objections do not hold.

It's Hard Enough Without AI: The Value of Diagnosis in Veterinary Practice

Mona F. Giersberg, Franck L.B. Meijboom

Recently developed smartphone apps promise to reliably measure in real-time whether a cat or a horse is in pain based on facial features (e.g. Feline Grimace Scale-, EPWA app). The use of such tools by caretakers or veterinarians can have direct consequences for the animal patient, particularly as animals lack self-reporting capacities and are subject to others' assessments of their condition. Risks for the animal that may lead to poor treatment include under-, over- or misdiagnosis by the app, bias towards certain breeds, ages or anatomical features, and lack of algorithmic transparency. In practice, these risks depend on the accuracy and validity of the app and are likely to decrease with technological progress. In our presentation, we will address a more hidden risk of AI in medical decision making, which may impact the treatment of a patient through shifts in veterinary professional identities. AI-based pain detection apps can partly or completely take over what has traditionally been the veterinarian's expertise: providing a diagnosis. Diagnosis in veterinary medicine is more than the basis for adequate treatment and good healthcare. It is also a social momentum for the veterinarian to build relationships and constitute professional power and authority. Maintaining this authority is important for the veterinarian to be able to enact their desired professional identity, particularly a patient-centred identity. We will argue that AI solutions may challenge and disrupt expertise-boundaries, alter the value of diagnosis, and ultimately may lead to identity confusion. Identity confusion in practice can mean that a veterinarian struggles to make decisions in the veterinarian-caretaker-animal-public expectation context or that they are more prone to adopting a caretaker-centred position. We will conclude with defining preconditions for the development and application of pain detection apps in veterinary practice so that they add to the value of diagnosis instead of eroding it.

The Synergy between AI and Biotechnology for Conservation: an Epistemic Justice Problem

Bernice Bovenkerk, Dominic Lenzi

In response to the current biodiversity crisis, the field of biodiversity conservation has seen two recent developments. Firstly, there is increasing consensus that halting biodiversity decline must be inclusive of the perspectives and knowledge of indigenous peoples and local groups. Secondly, biodiversity conservation has become increasingly dependent on technological interventions. A new combination of technologies used in biodiversity conservation is the synergy between AI and biotechnology. AI could fill in the gaps when genetic information is missing and tell us which genetic variants are most likely to produce the best adaptations to changing environmental pressures. An AI model could even 'custom design' DNA constructs, to be engineered into living organisms, and released into the environment.

We argue that this synergy challenges the recent development towards more inclusive biodiversity conservation. The so-called 'black box problem' raises epistemic and recognition justice issues in three specific contexts: (1) training of data, (2) valuation and decision-making process, and (3) dealing with risks and responsibilities. The biases in training data and the limited ability of available scientific data to represent plural perspectives on conservation challenge the meaningful inclusion of these perspectives in AI tools. Moreover, the worldviews and priorities of AI developers, and the financial interests backing them, are likely to play a decisive role in shaping and applying these tools in conservation. Finally, to a certain extent the inner workings of the algorithms are unknown even to their developers themselves, and subsequently we are not only dealing with risks, but also with unknown unknowns.

These features highlight how epistemic injustice is a plausible outcome of utilizing AI in conservation. We suggest that in order to address the tension between the two developments of inclusivity and technology employment it is essential to create inclusive processes to determine conservation priorities before (and whether) algorithms are designed.

Unlearning Human Bias: Teaching Language Models to Think Beyond Species

Christoph Krüger, Sam-Tucker Davis

This research presents a novel approach to teaching LLMs animal ethics through a combination of NGO-sourced data, predictive models, and synthetic dialogue.

Our methodology consists of three key phases: (1) Data collection from NGOs providing two distinct datasets: factual information about animal welfare and historical performance data showing which advocacy approaches have been most effective. (2) Development of an alignment prediction model trained through a combination of human feedback and synthetic human and non-human animal perspectives. (3) Model alignment, where we fine-tune Llama 3.2 using only synthetic dialogues that score highly on both ethical alignment and predicted real-world effectiveness. Llama 3.2 was chosen, as it is one of the state-of-the-art LLMs, which is open source, facilitating transparency and open access of the resulting model.

This study contributes three key innovations: an open-source fine-tuned LLM optimized for animal welfare discussions, a novel alignment prediction framework combining human and synthetic feedback, and a methodology for generating highly effective advocacy dialogues using NGO-sourced facts and performance data. Our results demonstrate that machines can be trained to advocate for animals by combining factual accuracy, ethical alignment, and empirically-validated persuasion strategies. This approach lays the foundation for AI systems that can effectively expand humanity's moral circle.

Our empirical evaluation shows substantial reduction in speciesist bias across the Animal Harm Assessment benchmarking while showing limited tradeoffs in performance on standard language generation benchmarks (LAMBDA, HellaSwag, WInoGrade and ARC-Challenge). These findings suggest that addressing speciesism may strengthen fundamental ethical reasoning capabilities in language models. We hypothesize that training models to extend moral consideration beyond arbitrary species boundaries enhances their capacity to recognize and resist other forms of group-based discrimination. This supports our central thesis that speciesism functions as a foundational bias that, when addressed, enables more consistent ethical reasoning across domains.

Teaching With and About AI

Preprint community: https://zenodo.org/communities/iacap-aisb-25-teachingai

In this symposium, we will take the following chapters and present the central arguments of each. The intention of the symposium is to get flow on the connections between chapters and to get audience feedback on missing chapters or useful links between them. The volume will start with academic works by scholars, and end with student-authored case studies of their learning and insights. We will not be presenting traditional papers here, instead, we will be trying to get feedback on the volume as a whole.

Workshop booklet: https://zenodo.org/records/15741641

We will present the arguments from the following papers:

- The Emperor's New Clothes: A Manifesto for Universities in an AI-Haunted World (Ballsun-Stanton, Khalid)
- Towards an Evidence Based Framework for AI in Philosophy Education (Parks)
- Teaching the Unknown: A Pedagogical Framework for Teaching With and About AI (Ballsun-Stanton, Torrington)
- Ethnographic insights into AI Pedagogy for Higher Education: A Qualitative Study. (Martinelli)

A dataset of assignments and class transcripts from a full semester of teaching the pragmatics of AI to humanities students.

Case studies by undergraduate students:

- AI's Applicability to the Legal Profession
- The Art of the Prompt: Lessons in Play, Bias, and Digital Creation
- Grappling with uncertainty in AI use
- Trust in our tools: a spectrum of AI utility
- Rethinking the Role of AI in University Learning

Interactive Workshop: Writing and Research with (and about) AI: using Claude and OpenAI's LLMs to Scaffold and Edit Papers

Presented as an extension of the Teaching with and About AI workshop (feel free to attend one or both sessions).

Demonstrating techniques from Ballsun-Stanton's Pragmatics and Ethics of AI classes, this interactive workshop will allow participants to explore and test ideas from the research submitted for the Teaching with and About AI volume. We will spend the last half hour workshopping AI-Enabled student assessments for the humanities using evidence from prior semesters teaching and assessments across the Faculty of Arts.

Topics covered:

In-class simulations

Prompting for effective scaffolding (ideation and outlining)

Prompting for copy-editing

Thoughts on how to use "agentic" (they're not) tools.

For a best experience, please register for console.anthropic.com ahead of time (requires phone number) as this will give you $5 of free credits to use as part of the workshop. (If everyone tries to register at once, AI providers tend to interpret it as spam and lock people out.)

The Epistemic Risks of AI Integration

The Limitation Game: Antrophomorphising and AI Testimony

Ian Robertson

How does the way we learn from AI systems structurally parallel with the way that we learn from human experts? More specifically, how tenable is it to adopt our most prominent accounts of testimonial knowledge as a model for understanding how we glean warrant from trusting the deliverances of AI systems? In this talk, I examine recent attempts to advance tenable accounts of 'AI testimony' and show them to be wanting. I begin by showing a series of candidate ways in which AI systems are utilised to generate recommendations and compare them to other epistemic environments where they are construed as yielding preliminary guides for further inquiry. Having done so, I show that recent attempts to ground AI testimony in anti-reductionist terms are unpromising. Ultimately, it is argued, there are good reasons to construe AI systems not as testifiers but to limit them to a kind of epistemic tool.

AI and the Heterogeneity of Pain

Hadeel Naeem

The increasing reliance on opaque AI risks erasing the heterogeneity inherent in some concepts. Pain, for instance, is a complex and multifaceted phenomenon. Nociceptive signals from the body interact with psychological, cultural, and contextual factors, creating an intrinsically subjective and varied experience. Many AI tools that assess pain focus predominantly on bodily and behavioural changes, such as facial expressions, vocalisations, or physiological metrics. By doing so, these tools risk defining pain as a homogeneous concept and overlooking its subjective and nuanced dimensions.

Varieties of Epistemic Risks in Emerging Technologies

Sascha Fink

Some technologies have the potential to make us worse at knowing (epistemic risk) under some conditions while also having the potential to make us better at knowing (epistemic opportunity) under other conditions. Currently, the debate in Anti-Risk Epistemology focuses primarily on forming false beliefs and on failing to form true beliefs. However, there could be many more kinds of epistemic risks at play especially due to the opacity of AI systems. Here, I will present several epistemic risks other than the two currently discussed in Anti-Risk Epistemology and illustrate them with concrete examples.

Epistemic Vices: How AI Shapes Attention, Imagination and Other (Intellectual) Virtues

Deb Marber

In this talk, I will focus on the epistemic risks our engagement with AI poses in terms not just of our ability to produce or acquire knowledge and other epistemic goods reliably, but also of the epistemic processes and habits it shapes in its users through promoting or hindering various virtues (cf., e.g., Vallor 2024; 2016, and Smith and Vickers 2024, Russo et al. 2024, Ohlhorst 2025). In particular, I will argue that our current engagement with AI tools facilitates epistemic vices such as impatience, laziness, intellectual cowardice and preference and attention patterns such as risk and uncertainty aversion which too often result in arrogance, stubbornness and other undesirable traits in knowers. I show how, in turn, these impact our ability to imagine, constituting intellectual vices of the imagination (the flipside of, e.g., Marber and Wilson 2024) with major implications for our epistemic resilience and a heightened epistemic risk.

The Values and Disruptive Capacities of AI Systems

Misalignment or Misuse? A Tradeoff

Max Hellrigel-Holderbaum

Creating systems that are aligned with our goals is seen as a leading approach to create safe and beneficial AI in both leading AI companies and the academic field of AI safety. We defend the view that misaligned AGI – future, generally intelligent (robotic) AI agents – poses catastrophic risks. At the same time, we support the view that aligned AGI creates a substantial risk of catastrophic misuse by humans. While both risks are severe and stand

in tension with one another, we show that – in principle – there is room for alignment approaches which do not increase misuse risk. We then investigate how the tradeoff between misalignment and misuse looks empirically for different technical approaches to AI alignment. Here, we argue that many current alignment techniques and foreseeable improvements thereof plausibly increase risks of catastrophic misuse. Since the impacts of AI depend on the social context, we close by discussing important social factors and suggest that to reduce the risk of a misuse catastrophe due to aligned AGI, techniques such as robustness, AI control methods and especially good governance seem essential.

AI Outsourcing and the Value of Autonomy

Eleonora Catena

The development of AI technologies has the potential to disrupt and transform human values (Danaher & Sætra, 2022; 2023; Danaher, 2024). Among them, autonomy is widely considered a key value for both individuals and societies. However, autonomy is also affected and shaped by technological development. Consequently, the interaction with AI technologies that mediate human decisions and actions has crucial implications for human autonomy (e.g., Laitinen & Sahlgren, 2021; Chiodo, 2022; Prunkl, 2024). This paper investigates whether and how AI outsourcing changes the value of human autonomy by affecting our exercise, understanding, and valuing of it. The first part clarifies the notions of outsourcing, control, and autonomy to analyse their interaction. By definition, AI outsourcing entails a trade-off between process control and outcome control (see di Nucci, 2020; Constantinescu, 2025): offloading tasks to AI technologies implies transferring control of the underlying processes for improved control over the pursued outcomes. This process maps onto two components of autonomy, intended as self-control: the control over one's internal processes (e.g., deliberation and decision-making) and the control over external outcomes (e.g., goals and life plans). Based on this analysis, the second part of the paper draws the implications for the value of human autonomy. The main suggestion is that AI outsourcing favours a re-prioritization concerning the understanding and valuing of autonomy. This change follows from the internal trade-off that favours one component of autonomy and disincentivizes the other. Moreover, AI outsourcing raises anew the key challenge to the value of autonomy: whether there is something intrinsically or instrumentally valuable in control of oneself. The concluding discussion sketches further implications, including risks and open questions, raised by AI outsourcing for the value of autonomy.

AI Value Alignment: From Rights to Capabilities

Ibifuro R. Jaja

Given that AI systems are increasingly deployed to make decisions with significant impacts on individuals and societies, there is a consensus that AI systems should be value-aligned. This push for value-aligned AI systems is to ensure that AI-driven decisions do not have harmful or unjust impacts. However, there is a disagreement about what values AI systems should be aligned with (Robinson, 2024; Floridi & Cowls, 2019). This disagreement stems from the nature of values themselves as context-dependent. What is considered morally valuable in one context may not be considered so in other cultures (Whittlestone et.al., 2019). Moreover, each perspective may be supported by well-developed and formulated arguments explaining why theirs are valid or why others are less so. One proposal put forward by scholars is to adopt the human rights framework as the guiding set of values for aligning AI systems (Gordon, 2023; Smuha, 2020; Yeung et al., 2020). The appeal of the human rights framework lies in its universal recognition (Smuha, 2020). While the degree to which it is upheld may vary across countries, it has been ratified by all member states of the United Nations. In this regard, the human rights framework can be considered the closest thing to a universally shared set of values (Nickel, 2007). In this talk, I argue that a human rights approach to AI value alignment is redundant. While it offers basic protections for human dignity, it does not address the full range of ethical, social, and existential impacts AI systems may have. Since there is a universal recognition of the aspects that it covers, their fulfillment should be assumed or enforced rather than requiring explicit reaffirmation. Consequently, the call for a human rights approach does not significantly advance the discussion of AI value alignment. I advance the capability approach developed by Amartya Sen and expanded by Nussbaum, which evaluates individual well-being based on real freedoms to achieve their goals, as a framework for AI value alignment.

Short-term or Long-term AI Ethics? A Dilemma for Fanatics

Vincent C. Müller

There seems to be a dilemma whether we should direct our efforts in AI ethics towards the problems that are visible today, or on the horizon (short-term), or towards extremely important problems for which we see significant risk of them occurring at some point (long-term). Some authors have argued that we should ignore the one or the other, calling short-termists "short-sighted", or calling long-termists "singularitarians". I will argue that this is a false dilemma. (1) While any rational agent will consider short- and long-term consequences, the supposed dilemma rests on the assumption that there is a difference between the two kinds of problems which

is significant enough to force an exclusion of one option – what I call the "significant difference view". (2) The only serious argument for this view is that the longer term involves an ethically truly new situation, a "singularity", and that this demands a version of ethical "fanaticism". (3) I will undermine the first component of that conjunction (in 2) by presenting a positive argument towards the view that the short- and long-term problems do not have such a "significant difference". (4) I explain how fanaticism implies meta-ethical fanaticism. (5) The conclusion is that the dilemma only occurs if one makes too many dubious assumptions. We should return to the "normal balance" of expected utility in AI ethics.

POSTER PRESENTATIONS

GenAI for Innovation: Framing Trust

Koen Bruynseels

Generative AI (GenAI) is increasingly used in Research and Innovation. GenAI models allow for capturing relations in large bodies of textual data and were therefore defined as "epistemic technologies". GenAI though is also applied in innovation, which requires a relation to the GenAI model that goes beyond mere epistemic reliance. Innovation with the use of GenAI implies the context of discovery, in which GenAI is not used primarily to make knowledge claims, but also to identify a subset of candidates within a large search space. This brings in the notion of trust. Trust, in the account of Annette Baier, requires the goodwill of a person to not take advantage of your vulnerability to harm "the goods or things one values or cares about". Trust in sociotechnical systems that use GenAI to support their innovations therefore relates to the question of whether these values are taken into consideration.

On the Establishment of Ethics in Autonomous Intelligent Systems

Aaron Joseph Butler

The aim of the project is to investigate whether there can be developed a viable account (i.e., a model) on the basis of which ethics can be established as an internal feature of autonomous intelligent systems (AIS). The research questions addressed are: Is there a viable model capable of establishing ethics as an internal feature of AISs? What are the principled reasons for it, and how would it work? The work herein is executed over three phases. First, the problematic is articulated; in so doing, the following intuitive question, namely: 'When behaviour of an autonomous intelligent system(s) leads to harm or death of someone, who is responsible?', is used as a stepping off point for the inquiry. Second, a practice-oriented conceptual framework and model, the union of which constitutes the theoretical basis whose viability will be assessed, is developed: expressing the insight that models are created out of one's total understanding of how the world works. Third, the theoretical basis is tested as a "proof of concept" against three key challenges to any account capable of establishing ethics as an internal feature of the relevant technical systems, namely: autonomy problem (2) opacity problem (3) enforceability problem.

Evaluating Compositionality in Large Language Models Through Natural Language to First-order Logic Translation

İbrahim Ethem Deveci

Transformer-based large language models have achieved significant success in natural language processing, but their ability to adhere the principle of compositionality without relying on explicit symbolic representations remains a point of contention in cognitive science and the philosophy of artificial intelligence. This work evaluates the capacity of large language models to translate natural language sentences into first-order logic expressions. This task requires handling the contextual flexibility of natural language while maintaining the formal rigor of logical expressions, with implications for the models' capacity for language understanding and their ability to adhere compositionality. We conduct experiments using multiple large language models, employing various techniques, including fine-tuning, zero-shot, and few-shot prompting. Our approach provides a comprehensive evaluation framework, comparing model performance across different architectures and parameter sizes. Beyond traditional evaluation metrics, we emphasize the need for task-specific metrics that assess the properties of first-order logic expressions, such as well-formedness, the distinction between form and content, and the equivalence of different logical expressions conveying the same meaning. By offering a nuanced evaluation, this work informs ongoing discussions in cognitive science and philosophy of artificial intelligence, while advancing research on natural language processing and semantic parsing.

Data Scientists and Society: Fostering Critical Thinking and Societal Engagement

Heike Felzmann

In this paper we present our initial results of a project on engaging data scientists in critical reflection on the meaning and social impact of their research and professional practice. We were looking to (i) understand researchers' and practitioners' perceptions of societal and cultural issues relating to data science and AI and (ii) build researchers' capacity for reflection on such issues. The project combined interviews with data science researchers and practitioners with a set of interventions using "communities of inquiry" (CoIs) with researchers in an Irish data science centre. CoIs consist in facilitated in-depth group reflection on substantive questions that are centred around fundamental concepts. These included both a series of once-off reflective events within the institute and a pilot training of early career researchers to enable them to use CoIs themselves to engage members of the public in dialogue about their research. We will describe our approach, report initial

results, and reflect on our experiences as humanities researchers within the data science context.

What's the Problem with Anthropomorphising AI-driven Systems?

Giles Howdle

According to 'a widespread view' (Coghlan, 2024), our anthropomorphic way of thinking and talking about AI-driven systems is a mistake. I distinguish two interpretations of the supposed anthropomorphic mistake, metaphysical and pragmatic. I object to the metaphysical interpretation and develop the pragmatic interpretation.

On the metaphysical interpretation, the mistake is that our thoughts and utterances carry a commitment to ontological falsehoods, e.g. to the existence of (non-existent) artificial minds. I provide two objections. First, we may be using non-literal or metaphorical anthropomorphic ascriptions that do not carry an ontological commitment. Second, if we are committing ourselves to ontological falsehoods when talking and thinking about AI, then we would also doing so when we anthropomorphise corporations and thermostats. But this is implausible.

These objections motivate an alternative, pragmatic interpretation of the anthropomorphic mistake. It is not that our AI-related thought and talk fail to correspond with reality; rather, we are adopting a way of thinking and speaking that can get us into trouble. The mistake is that thinking and talking anthropomorphically about AI-driven systems leads to (vulnerability to) predictive error, which can have negative downstream consequences, including leading us to make poor inferences.

Pythagorean Path, Ontological Anxiety and Cold Death of Bitcoin

Daniel Hromada, Harashi Namztohoto

Some time ago, at the IACAP 2013 conference, a decentralized monetary innovation known as Bitcoin has been labeled as a sort of new "religion" in a world marked by Nietzsche's "death of God." This paper revisits that claim, exploring Bitcoin's ultimate trajectory as a temporary proxy in humanity's transfer of power to machines, grounded in Nietzsche's "transvaluation of all values."

Central to this argument is the "Pythagorean path," which highlights Bitcoin's reliance on the Elliptic Curve Discrete Logarithm Problem (ECDLP). The security of Bitcoin rests on the irreversibility of scalar multiplication on elliptic curves, but this foundation may not be immune to future mathematical breakthroughs.

The paper also addresses the "ontologic anxiety" inherent in Bitcoin's fragile survival and critiques the misplaced confidence in its longevity, often justified by the Lindy effect. Finally, it considers Bitcoin's vulnerability to non-deterministic events, positing that minor disruptions could cascade into system's "cold death."

Posthuman Creative Styling – a Philosophy and Model for Representing the Actions of Creative Individuals When Generating Creative Writing.

Christopher Mart

This presentation is about a joint discipline philosophy called 'posthuman creative styling'. It presents a model for creative styling actions applicable to both machine writers and people. The format for the model is based on a conceptual space framework which has been developed using a terminology borrowed from *computer science* and object-oriented programming (OOP). The method used was to allow *creative writing* practice to inform the model's outcome. It did this by a process of discovery for the three types of creative action: exploration, transformation and combination of conceptual spaces and developed a procedural notation for each. By encoding style rules into formulae, ninety-nine pieces of creative writing were generated by a creative individual, with notes made for how the qualities of the writing's style had their value calculated.

The qualities covered many types of conceptual space for the different and varied style actions that creative writers use. Using mnemonics to simplify the notation, the model demonstrates how the processes of all creative individuals can be described. The presentation will show how its way of describing style as a kind of atomic entity is a practical way forward for researchers interested in encoding style creativity.

The Virtuous Machine: Extended Cognition as Scaffolding for Artificial Moral Development

Justas Petronis

When artificial intelligence systems already increasingly shape our decision-making processes, this presentation presents a radical reconceptualization of artificial moral development. By synthesizing Clark and Chalmers' extended mind thesis with Vallor's technomoral virtue ethics, presentation proposes a framework that moves beyond the limitations of current principlism-based approaches to AI ethics. Rather than treating AI systems as independent moral agents, we position them as components of extended moral cognitive systems, inextricably linked with human moral agents. The framework leverages predictive processing architectures and the "controlled hallucination" model of cognition to establish a foundation for genuine moral perception and judgment in artificial systems. Through the lens of four core virtues - justice, honesty, responsibility, and care - we demonstrate how moral capabilities can be developed through extended cognitive scaffolding and dynamic interaction with human moral agents. This approach

addresses the longstanding "ELIZA problem" while offering practical implementation strategies for virtue-based AI systems. The research contributes to multiple domains, including AI ethics, cognitive science, and embodied cognition, with implications extending from theoretical frameworks to practical AI development and deployment strategies.

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

Krzysztof Sołoducha

The objective of my talk will be to present first results of ongoing research project which pursues to find an answer to the question of how the emerging technology of human assisting embodied robots can be equipped with a system of simulating the attitudes and moral values of its users using contemporary methods of digital humanities. Therefore we started to build a simple and effective system for identifying of the ethical preferences of users of human assisting social machines – recognising the explicit and implicit normativity influencing their ethical decisions. In the next step, these identified implicit and explicit normativities should be implemented into a system of their digital simulation. The platform for doing so will be multi-agent AI technology, which is regarded as a so-called complex (compound) artificial intelligence system.