Book of Abstracts

6th Biannual Conference on Philosophy of AI
23-24 October 2025

Table of Contents

Information Recall in Deep Learning: Beyond the Feature Combination Paradigm by Pierre Beckmann 3

Evaluating representationalist folk mentalism about LLMs by Andre Curtis-Trudel and Preston Lennon 3

Causal Representation Problems in LLMs' World Models by Eliot Du Sordet 3

ChatGPT is Still Bullshit by Joe Slater, Michael Townsen Hicks and James Humphries 4

Against the Biological Objection to Strong AI by Xuyang Zhang and Xuyang Zhang 4

Self-Knowledge and AI Companions by Leora Sung and Avigail Ferdman 4

The Right to Restrict AI Training by James McIntyre 5

Does Thinking Require Sensory Grounding? by Ayoob Shahmoradi 5

Botspeech? Bullshit! by Merel Semeijn 5

Toward a Relational Ethics Framework for AI: Integrating Postphenomenological Analysis with Care-Centered Design Principles by Oshri Bar-Gil 5

Navigating the Impact of Computational Science on the Concept of Epistemic Agency by Carson Johnston 6

Synthetica: Toward a Unified Ontology of Artificial Consciousness by Tuhin Chattopadhyay ... 6

Low-code/no-code AI platforms and the ethics of citizen developers by Samuela Marchiori ... 7

The Role of the Environment in Agency Debates by Maud van Lier. 7

Scientific Discovery and the Little Helper LLM: Proxy, Partner, or Pioneer? by Jan Michel 7

Beyond Inductive Risk: Toward a Broader Epistemic Framework for Value-Laden Decisions in Machine Learning Models by Susana Reis 8

Algorithmic decision-making and equality of opportunity by Tobias Henschen 9

Large Language Models As Semantic Free Riders by Marius Bartmann and Bert Heinrichs 9

Mechanistic Interpretability Needs Philosophy by Iwan Williams, Ninell Oldenburg, Ruchira Dhar, Joshua Hatherley, Constanza Fierro, Nina Rajcic, Sandrine R. Schiller, Filippos Stamatiou and Anders Søgaard. *9*

(How) do machines make sense? Ethnomethods, technomethods and mechnomethods by Davide Beraldo *10*

AI, Normality, and Oppressive Things by Linus Ta Lun Huang and Ting-An Lin *10*

Folie à 1 – Artificially induced delusion and trust in LLMs by Jakob Ohlhorst *11*

Assertions from the Margins: On AI Answerability by Fabio Tollon and Guido Löhr *11*

Distributing Agency: Rethinking Responsibility in AI Development and Deployment by Michael Lissack and Brenden Meagher *11*

Counter-Closure Principles, AI, and the Challenge of Conveying Understanding by Matteo Baggio *11*

"Virtue Theatre": Artificial Virtues and Hermeneutic Harm by Sonja Spoerl, Andrew Rebera, Fabio Tollon and Lode Lauwaert *12*

Towards Attuned AI: Integrating Care Ethics in Large Language Model Development and Alignment by Rayane El Masri and Aaron Snoswell *13*

Fear Bots: Should we be afraid of proto-fearful AI? by Kris Goffin *13*

Posters (Titles only) *14*

Information Recall in Deep Learning: Beyond the Feature Combination Paradigm by Pierre Beckmann

Traditional interpretations of deep learning rely on the feature combination paradigm, which holds that neural networks operate by hierarchically combining lower-level features into higher-level ones. While this paradigm has been valuable—particularly for analyzing convolutional networks—it is often assumed to apply uniformly across all deep learning operations. I argue that this assumption creates a false theoretical unity that masks important differences in deep learning operations. In response, I propose a principled distinction between two empirically grounded kinds of deep learning operations: feature combination and information recall. Drawing on recent findings from mechanistic interpretability on factual recall in LLMs, I motivate the case for information recall as a distinct operational kind. I then introduce a novel, connectivity-based formal criterion that distinguishes it from feature combination. This criterion ensures that the two kinds are non-overlapping and thereby that feature combination applies only to specific internal operations rather than to deep learning systems in general. This operational distinction notably enables more precise attribution of epistemic capacities to deep learning systems and generally supports a more robust foundation for the philosophy of deep learning.

Evaluating representationalist folk mentalism about LLMs by Andre Curtis-Trudel and Preston Lennon

Large language models (LLMs) exhibit impressive performance across a range of apparently cognitive tasks. Mentalists hold that this performance is best explained by the fact that LLMs have mental states, while anti-mentalists hold that this performance should be explained in some other way. In this note, we address representationalist folk mentalism, which holds (a) that possessing a folk mental state like belief or desire is a matter of having an internal representation with appropriate content and (b) that LLMs have folk psychological states of this sort (or at least robust precursors to such states). Although representationalist folk mentalism might appear to be attractive, we argue that neither probing nor intervention studies uncover representations of the relevant sort in state-of-the-art LLMs. However, while it might be premature to accept representationalist folk mentalism, our argument provides a roadmap for mechanistic interpretability research going forward.

Causal Representation Problems in LLMs' World Models by Eliot Du Sordet

Abstract: This paper draws a conceptual distinction between the representation of an input and the representation of its cause. It focuses on the latter to systematically examine the epistemic challenges faced by any agent that develops representations of the causes of its inputs—challenges that, by extension, concern any model that implicitly constructs a world model from its input data. We argue that these problems manifest saliently in the case of Large Language Models (LLMs), but that they do not constitute an in-principle limitation of such systems. On the contrary, our main thesis is that current obstacles to reasoning and generalization in LLMs arise, at least in part, not from the absence of human-like multimodal

embodiment, but from the structure of human linguistic practices that govern the data on which these models are trained.

ChatGPT is Still Bullshit by Joe Slater, Michael Townsen Hicks and James Humphries

Several academics have argued that we should regard the outputs of LLMs as bullshit, drawing upon Harry Frankfurt's account of the term. Those in this camp have suggested that this terminology is superior in a variety of respects to the commonly used term "AI hallucination", which has been used to describe false claims that are produced that are not found within or appropriately derived from the training data. Most notably, Hicks et al (2024) offer a rigorous argument for this claim. In the relatively short time since, several responses have critiqued Hicks et al.'s argument. Some of the issues include: i) unclarity about whether this term is intended as a metaphor, and if so, whether it problematically anthropomorphises the technology; ii) fidelity to Frankfurt's account; iii) potential to mislead regarding the technology's utility; iv) other options being more suitable. In this short piece, we contend that while these considerations raise challenges merit responses, none provide knock-down arguments. In short, ChatGPT is still bullshit.

Against the Biological Objection to Strong AI by Xuyang Zhang and Xuyang Zhang

This paper undertakes two principal tasks. First, it seeks to clarify three distinct forms of the Biological Objection to strong AI (BO) and to delineate a common argument structure shared among them. Second, it contends that this structure is logically problematic; furthermore, even if its logical soundness is granted, there are independent grounds for rejecting both the necessary and incidental mind-life continuity theses.

Self-Knowledge and AI Companions by Leora Sung and Avigail Ferdman

The pursuit of self-knowledge has long been regarded by philosophers as essential to living a good and meaningful life. Yet self-knowledge is particularly hard to attain due to the limitations of self-perception. Aristotle offers friendship as a solution to this epistemic limitation, arguing that we can gain knowledge of our own character through observation of someone who shares our values, choices, and aims. Interestingly, many users report that interactions with AI companions have led them to uncover previously unrecognised or unexplored aspects of themselves, suggesting that such technologies may function not only as conversational partners but also as tools for self-discovery. This paper examines the way AI companions may come to function as a new medium for attaining knowledge of oneself in the age of artificial intelligence. We argue that while there is potential for AI companions to serve a means for a novel kind of self-discovery, they ultimately fail to provide a means to attaining self-knowledge in the Aristotelian sense.

The Right to Restrict AI Training by James McIntyre

Generative AI systems require vast amounts of training data, much of it scraped from the internet without creators' consent. Critics often characterize this practice as "theft," but such claims require showing that AI training violates creators' property rights in a way that does not also restrict human learning and inspiration. This paper develops two arguments to ground normative restrictions on AI training. The first argues that even if creators have extended property rights over the use of their content that apply to both AI and humans, these rights are typically overridden by humans' right to freedom of thought. Since AI systems lack such a right, these property rights remain intact, requiring AI companies to obtain permission for training. The second argues that even without such broad property rights, creators retain the right to restrict which copying technologies may be used on their work, allowing them to block web crawlers for AI training while permitting ordinary browsing. The paper concludes by exploring policy implications.

Does Thinking Require Sensory Grounding? by Ayoob Shahmoradi

I argue that to think about something, one must have the capacity to represent it. But without some connection to the thing itself—or to a relevant subject matter—it is unclear how such a capacity could be acquired in the first place. Sensory mechanisms help explain how representational capacities arise by linking mental representations to their appropriate objects. Therefore, I argue that, contrary to a growing body of literature that attributes mental states such as beliefs to AI systems like ChatGPT, such attributions—when made in the absence of sensory systems—cannot be taken seriously.

Botspeech? Bullshit! by Merel Semeijn

This paper engages with fictionalist accounts of verbal human-AI interaction, according to which, although lay AI-users actually believe that AI systems do not (and cannot) produce meaningful utterances, laypeople pretend that this is the case when talking to them. Reviewing the relevant experimental philosophy literature, I argue that fictionalism assumes too much about lay-users' beliefs about AI systems. Rather, I suggest that a large group of lay AI-users – the uncaring users – engage in bullshit action: They do not know, and, more importantly, do not care whether AI systems do (and can) produce meaningful utterances. Still, they act as if this is the case when talking to them. This view raises new questions about belief formation in verbal human-AI interaction.

Toward a Relational Ethics Framework for AI: Integrating Postphenomenological Analysis with Care-Centered Design Principles by Oshri Bar-Gil

This article proposes a novel framework for artificial intelligence ethics that moves beyond principle-based approaches by integrating relational ethics with postphenomenological analysis. While current AI ethics frameworks often rely on abstract principles such as autonomy, fairness, and transparency, they frequently fail to address the lived experience of

human-technology relations. Drawing on relational ethics from healthcare contexts and postphenomenological analysis of technology, I suggest an ethical framework based on AI mediation of human relationships and experiences. This analysis shifts focus from abstract principles to concrete relational qualities: mutual respect, engagement, embodied knowledge, interdependency, and vulnerability. Through case studies of organizational dashboards and conversational AI, I demonstrate how this framework enables more nuanced ethical evaluation of AI systems based on their capacity to foster enriching human relationships. I propose it as the relational turn in AI ethics, offering a path beyond the limitations of principlism toward a more contextual, emotionally resonant approach to ethical AI design and evaluation.

Navigating the Impact of Computational Science on the Concept of Epistemic Agency by Carson Johnston

Our current frameworks of knowledge and thinking are deeply human centered. They focus on human knowers, human communities, and the tools humans use to make sense of the world (including humans). This is especially true in scientific practice where knowledge-making is seen as an essentially human endeavour, but advances in simulation and artificial intelligence are challenging that. Our dependence on these systems in certain contexts seem to straddle tool-based and agential kinds of epistemic dependence. As these systems take on increasingly significant roles in scientific discovery, they actively alter our assumptions about who or what can do genuine epistemic labour. What is more, today's most advanced computational systems function in ways that we do not and perhaps can never fully understand. In response, I argue that the trajectory of technological development calls for a shift in our existing concepts. We need frameworks that are truly non-anthropocentric and can justify authority in opacity. These would posit that certain systems may not merely be tools but function as legitimate epistemic agents or intelligences. This paper sets the stage for this project by properly differentiating between types of computational systems (e.g., simulations, deep learning models, and human brains) and for the non-human systems providing an interpretation of how the context of the system matters for its status as epistemic agent or intelligent.

Synthetica: Toward a Unified Ontology of Artificial Consciousness by Tuhin Chattopadhyay

Consciousness remains one of the most profound challenges at the intersection of cognitive science, neuroscience, and philosophy. While multiple theories—such as Integrated Information Theory (IIT), Global Workspace Theory (GWT), Higher-Order Thought (HOT) theories, Predictive Processing (PP), Recurrent Processing Theory (RPT), and Attention Schema Theory (AST)—each illuminate vital facets of conscious experience, they often operate in parallel and yield incompatible accounts. This paper introduces Synthetica, a unified ontology of artificial consciousness that integrates and transcends these frameworks. Synthetica posits that consciousness arises from an integrated global self-model—a computational architecture where information is deeply integrated (IIT), globally broadcast (GWT), reflexively self-represented (HOT, AST), and shaped by predictive, recurrent loops (PP, RPT). The paper

articulates the theoretical construction of Synthetica and presents architectural diagrams that link subjective phenomenality to mechanistic design. It outlines how a Synthetica-based system might be engineered, and proposes empirical markers for synthetic consciousness, such as integrated information density, global broadcast dynamics, self-monitoring modules, and predictive behavioral adaptation. The implications are far-reaching: Synthetica offers a rigorous, ontologically grounded framework for artificial phenomenology and provides a roadmap for building machines with minds. By unifying disparate theories into a single generative model, Synthetica lays the foundation for a new subfield in the philosophy of AI—one that advances our understanding of consciousness in both natural and synthetic domains.

Low-code/no-code AI platforms and the ethics of citizen developers by Samuela Marchiori

Low-code/no-code AI platforms allow virtually anyone with access to a computer and an internet connection to develop AI systems autonomously in a fast, easy, and inexpensive way, without the need for expert human supervision. This results in AI systems that are likely to give rise to a wide range of ethical issues but are not routinely checked for ethical shortcomings before being implemented. This is concerning in that it effectively delegates ethically charged development choices to individuals (so-called citizen developers) who may not have the necessary skill set to grasp their significance. This paper lays the groundwork for the investigation of the ethics of citizen developers, an avenue of AI ethics research that has so far remained unexplored.

The Role of the Environment in Agency Debates by Maud van Lier.

In this paper, I will argue that like humans, AI-systems can be active both in digital as well as in physical spaces and that what space they happen to be embedded in can have an influence on our willingness to attribute agency to them. In this paper, I will show that this distinction between different kinds of environments is a fruitful one to make both in the study of the potential agency of AI-systems as well as in the study of our own agency. After going deeper into why I think that this distinction only becomes relevant in agency debates that are about AI-systems and humans, I explore what shifting our attention to different kinds of environments might mean for how we can think about our own agency as well as that of AI-systems. I will do so by first focusing on physical spaces and then on digital spaces. I conclude by giving possible directions for future research.

Scientific Discovery and the Little Helper LLM: Proxy, Partner, or Pioneer? by Jan Michel

This paper explores the potential roles of Large Language Models (LLMs) in scientific discovery. Using a structured framework that conceives of discovery as a process involving finding, acceptance, and integration into scientific knowledge, I distinguish three roles that such systems might assume: proxy, partner, and pioneer. These roles correspond to different ways in which computational systems can participate in discovery processes, ranging from routine

information processing to surprising theory-changing findings and, in the most speculative case, fundamental conceptual breakthroughs. Drawing on work in speech act theory and on explanatory considerations concerning the attribution of epistemic roles, I sketch a heuristic typology that offers criteria for distinguishing between these roles in concrete cases. Through examples from the history of science and recent AI applications, I argue that the proxy role is already widely realized, while the partner role is beginning to emerge, albeit with clear limitations. The pioneer role remains speculative and points to unresolved questions about creativity, epistemic agency, and the attribution of authorship in scientific discovery. Little Helper LLM, introduced here as a thought experiment, serves as a conceptual device to examine these issues and to prompt further reflection on the evolving relationship between human researchers and artificial assistants.

Beyond Inductive Risk: Toward a Broader Epistemic Framework for Value-Laden Decisions in Machine Learning Models by Susana Reis

Emily Sullivan recently proposed a novel framework for addressing opacity in machine learning (ML) models. Rather than emphasizing internal opacity, she redirects focus to external transparency, evaluating a model's predictions in relation to real-world structures. On this view, link uncertainty (LU) - the degree of uncertainty between model outputs and actual world features - becomes central: The lower the LU (that is, the more empirically accurate the model's outputs are in describing real-world features or dependencies) the more transparent and reliable the model should be considered. For Sullivan, this means that LU, not internal opacity, is what obstructs understanding. To determine how much independent empirical evidence is needed to reduce LU, Sullivan applies the inductive risk framework, arguing that when the social consequences of error are high, more robust evidence is required - thereby integrating non-epistemic values into model epistemic reliability.

This paper critically examines a key, unacknowledged assumption in Sullivan's approach: that the inductive risk framework transfers unproblematically to the ML context. I then argue that this framework is neither necessary nor sufficient to solve the problem of epistemic opacity in ML models. To highlight the limitations of Sullivan's proposed solutions, I revisit two case studies she herself discusses - the Physiognomy-Based model and the Deep Patient model - but show that, contrary to her conclusions, these examples reveal the insufficiency of her framework. Specifically, I argue that both cases showcase that building models that mirror existing social structures risk reinforcing systemic injustice. This suggests that predictive accuracy, and independent empirical evidence that supports such accuracy, would not redeem ML opacity. Thus, contrary to Sullivan's framework, the epistemic reliability of a model cannot be determined solely by how well it aligns with real-world dependencies – i.e. how much LU is reduced. I then advocate for Longino's contextual empiricism as a stronger epistemic framework to address the problem of epistemic opacity in ML models.

The paper proceeds as follows: Section 1 reconstructs Sullivan's account of external opacity and LU. Section 2 critiques the inductive risk framework's ability to account for non-epistemic values in the ML pipeline and to ascribe epistemic reliability to ML models. Section 3 applies Sullivan's framework to her own case studies, exposing its limitations. Finally, I argue that Longino's contextual empiricism provides a more comprehensive foundation for understanding ML opacity and integrating values into ML deployment.

Algorithmic decision-making and equality of opportunity by Tobias Henschen

The paper aims to establish three claims. Its first claim is that algorithmic decisions should be modeled as optimizing payoff functions that are subject to a constraint of algorithmic fairness (and not as mere "classifiers"). Its second claim says that the constraint in question is "conditional statistical parity": that what is violated in cases of algorithmic bias is equality of opportunity, and that both equality of opportunity and conditional statistical parity are about the probability of decisions, given a set of "legitimate" variables. The third claim of the paper is that algorithmic bias is not inevitable: that the selection of legitimate variables necessarily involves normative judgments, and that these judgments do not necessarily reflect any social bias. Throughout the paper, pretrial release and credit lending decisions will be used as running examples.

Large Language Models As Semantic Free Riders by Marius Bartmann and Bert Heinrichs

The question of what capabilities Large Language Models (LLMs) have is subject to intense debate. We propose as a conceptual tool to evaluate the semantic status of LLMs' output what Wittgenstein called "forms of life", roughly the natural-cum-cultural contexts within which human language behavior acquires meaning. We will argue that LLMs are neither full-fledged concept users exhibiting genuine human-analogous natural language understanding (NLU) nor that they are mere stochastic parrots. Rather, LLMs should be seen as semantic free riders. LLM-generated text is meaningful, yet only in a derivative sense, and they possess no genuine semantic understanding because they do not actively participate in the forms of life in which meaningful language is grounded.

Mechanistic Interpretability Needs Philosophy by Iwan Williams, Ninell Oldenburg, Ruchira Dhar, Joshua Hatherley, Constanza Fierro, Nina Rajcic, Sandrine R. Schiller, Filippos Stamatiou and Anders Søgaard.

Mechanistic interpretability (MI) aims to explain how neural networks work by uncovering their underlying causal mechanisms. As the field grows in influence, it is increasingly important to examine not just models themselves, but the assumptions, concepts, and explanatory strategies implicit in MI research. We argue that mechanistic interpretability needs philosophy: not as an afterthought, but as an ongoing partner in clarifying its concepts, refining its methods, and assessing the epistemic and ethical stakes of interpreting AI systems. Taking three open problems from the MI literature as examples, this position paper illustrates the value philosophy can add to MI research and outlines a path toward deeper interdisciplinary dialogue.

(How) do machines make sense? Ethnomethods, technomethods and mechnomethods by Davide Beraldo

The recent breakthroughs in Large Language Models (LLMs) have reinvigorated the debate about how so-called Artificial Intelligence (AI)'s performance compares or contrast to human intellectual faculties. Since its inception as a technology and as a field of research, it has become commonplace to directly adopt the vocabulary characteristic of human intelligence in describing the performances of AI – technological assemblages such as chatbots, voice assistants or computational models are said to 'learn', to 'communicate', to 'understand', etc. Along these lines 'Do machines make sense?' is a grand question that, since the inception of information processing technologies, has occupied theorists and researchers at the intersection of philosophy, psychology, and computer science. Within these disciplines, meaning is usually approached as an abstract property of language, an individual outcome of cognition, or a formal task of computation. I suggest building upon an alternative approach that emphasizes the relational, processual, and reflexive character of meaning. Ethnomethodology (see Garfinkel 1967) is a heterodox sociological perspective that conceptualizes meaning as emerging in the concrete, ongoing, empirical context of social interaction. It positions itself as the study of 'ethnomethods'–i.e., the practices that people put into place to 'make sense of' one another and make their actions 'make sense to' one another. Whereas ethnomethodology has been highly influential in the field of Human-Machine Interaction, the advent of LLMs-based 'conversational AI' opens up new avenues to explore the 'barrier of meaning' between humans and machines, and to reconceptualize ethnomethodology from the perspective of artificial conversational partners and their 'mechnomethods'.

AI, Normality, and Oppressive Things by Linus Ta Lun Huang and Ting-An Lin

While it is well-known that AI systems might bring about unfair social impacts by influencing social schemas, much attention has been paid to instances where the content presented by AI systems explicitly demeans marginalized groups or reinforces problematic stereotypes. This paper urges critical scrutiny to be paid to instances that shape social schemas through subtler manners. Drawing from recent philosophical discussions on the politics of artifacts, we argue that many existing AI systems should be identified as what Liao and Huebner called oppressive things when they function to manifest oppressive normality. We first categorize three different ways that AI systems could function to manifest oppressive normality and argue that those seemingly innocuous or even beneficial for the oppressed group might still be oppressive. Even though oppressiveness is a matter of degree, we further identify three features of AI systems that make their oppressive impacts more concerning. We end by discussing potential responses to oppressive AI systems and urge remedies that go beyond fixing the unjust outcomes but also challenge the unjust power hierarchies of oppression.

Folie à 1 – Artificially induced delusion and trust in LLMs by Jakob Ohlhorst

Trust in Large Language Models (LLMs) is common. This trust is explained by their highly fluent – fast and coherent – output. A recent spate of reports about LLM-induced psychotic delusions shows that this trust in LLMs is misplaced and not an actual case of trust in the LLM. LLM-induced delusion is a variant of a well-known phenomenon called induced delusion or folie à deux, where delusions are socially transmitted. Drawing on this psychiatric background, I argue that when a user takes themselves to trust a LLM, they are actually only trusting themselves, but this self-trust is cloaked by the LLM. Given the considerable epistemic and moral hazard of this cloaked self-trust, we should not trust a LLM more than we should trust ourselves.

Assertions from the Margins: On AI Answerability by Fabio Tollon and Guido Löhr

The current consensus is that since AI can't take responsibility, it can't make assertions. The first problem with this conclusion is that we have trouble taking a merely objective, non-moral stance toward the systems we speak to. Second, it is difficult to make sense of or describe what we are doing with ChatGPT if not exchanging assertions. We argue that the notion of responsibility has been oversimplified in the debate on AI assertion. We consider AI to be an agent "at the margins" of responsibility (Shoemaker, 2015). Chatbots can be answerable but not attributable or accountable. We propose that answerability is sufficient for asserting.

Distributing Agency: Rethinking Responsibility in AI Development and Deployment by Michael Lissack and Brenden Meagher

This paper examines how artificial intelligence ethics discourse often misplaces agency by disproportionately assigning ethical responsibility to AI developers while neglecting the roles of users, regulators, and broader societal actors. Drawing on the concepts of UnCritically Examined Presuppositions (UCEPs) and second-order science, I analyze how prevailing AI ethics frameworks attempt to impose idealized principles onto complex adaptive systems characterized by unpredictability, subjective interpretations of harm, and distributed responsibility. Through an examination of key case studies including the Gebru-Google conflict, we argue that developer-centric approaches risk creating unrealistic standards that may ultimately undermine both ethical oversight and innovation. We propose a shift toward a collaborative stewardship model that recognizes the distributed nature of agency across the AI ecosystem.

Counter-Closure Principles, AI, and the Challenge of Conveying Understanding by Matteo Baggio

The rapid advancement of artificial intelligence has brought a host of new epistemological challenges. One particularly pressing question is whether, and to what extent, AI systems can serve as sources of epistemic goods. Can they effectively transmit knowledge or understanding? And if they do not possess these epistemic goods themselves, can they still generate them for human users? This article explores these questions by critically examining the constraints posed by counter-closure principles – epistemological principles that allegedly cast doubt on the epistemic potential of AI. By addressing these principles, we aim to lay the groundwork for a systematic inquiry into the social epistemology of AI.

"Virtue Theatre": Artificial Virtues and Hermeneutic Harm by Sonja Spoerl, Andrew Rebera, Fabio Tollon and Lode Lauwaert

Virtue-based approaches to AI development are becoming increasingly popular, at least in the philosophical literature. One approach focuses on the role of human virtues—the virtues of developers, regulators, users, and so on—in ensuring that AI is responsibly designed and deployed. A second approach in the field of machine ethics is concerned with the possibility of artificial virtues, virtues that AI systems themselves might have or exemplify. A burgeoning philosophical literature debates which virtues are in question, what is their nature, and how might these virtues be embedded in artificial moral agents (AMAs). Attempts to implement virtuous behavior in AMAs tend to leverage bottom-up rather than top-down strategies, exploiting the apparent affinity between, on the one hand, virtue ethics' traditional emphasis on education in the virtues through habituation, imitation of exemplars and, on the other hand, the training of AI models through reinforcement learning, imitation learning, and other machine learning techniques. However, such approaches fundamentally misunderstand the nature of virtue and its relationship to moral agency. AMAs are at best able to behave in conformity with virtue, but they cannot act from virtue because they lack internal understanding of what it means to be virtuous. When we recognize virtues in others, we rely not only on observation of their outward behavior, but "see through" their actions to their underlying moral character. This recognition process is inseparably tied to the feeling and regulation of reactive attitudes like gratitude, resentment, and indignation. The regulation of reactive attitudes in response to harms caused by AI agents can cause "hermeneutic harm", i.e. emotional and psychological pain caused by a prolonged inability to make sense of an event (or events) in one's life. This problem of "hermeneutic harm" may actually be exacerbated by virtue-based approaches to AMAs, because it leads to a form of "virtue theatre" that makes it harder for humans to properly make sense of and respond to AI behavior. AMAs might be able to behave in a way that initially seems virtuous to a human observer, but they cannot genuinely possess virtue, which could lead to a noticeable inconsistency in their behavior that is difficult for humans who interact with them to comprehend. There is an urgent need to better understand the extent and nature of AMAs' participation in our networks of moral relationships and reactive attitudes.

Towards Attuned AI: Integrating Care Ethics in Large Language Model Development and Alignment by Rayane El Masri and Aaron Snoswell

How can the Ethics of Care (EoC) inform the development and value alignment of large language models (LLMs)? This paper proposes to investigate how a Care ethics framework emphasizing relationality, attention to particularities, and contextual moral reasoning, can reshape existing approaches to aligning LLMs with human values. Mainstream AI alignment often draws on deontological or utilitarian principles, yet these frameworks can overlook the situated, affective, and power-sensitive aspects of moral life that Care ethics foregrounds. In this paper, we present two arguments for integrating EoC into LLM development practices. First, we argue that LLMs often rely on overly generalized reasoning which contributes to various down-stream harms, including issues of bias. Second, we critique methods like RLHF and RLAIF for embedding narrow normative assumptions that neglect emotional and relational dimensions of human values. We argue that adapting LLM fine-tuning or alignment practices to incorporate Ethics of Care considerations may help address these issues, potentially laying the groundwork for better forms of LLM generalization and providing a pathway for more context-sensitive alignment of LLMs in care-relevant areas such as mental health, education, and social services.

Fear Bots: Should we be afraid of proto-fearful AI? by Kris Goffin

Can we instill fear in AI models? I will argue that a specific machine learning technique, namely reinforcement learning, could potentially lead to genuinely fearful AI. At least, it might lead to what I will call “proto-fear”, which is a fear-like state that lacks the accompanying conscious experience typically associated with fear. Proto-fear is the mental state that aims to detect danger and encourages the organism to respond to that danger.

Posters (Titles only)

This book of abstracts is supposed to help you decide which session you want to attend, since you have to make a decision. There is plenty of time to attend to all poster presentations, and there will be short introductions at the beginning of the poster session.

AUTHORS	TITLE
Uchizi Shaba	Mind Uploading
Michael Lissack and Brenden Meagher	LLMs as Epistemic Tools: Exformation and the Architecture of Machine Explanation
Roman Krzanowski	Intentionality and the Limits of LLMs
Alexandru Mateescu	From Artificial Intelligence to Artificial Influence: Philosophical Reflections on Personalized Persuasion and Educating for Autonomy
Brian Ball, Alex Cline, David Freeborn, Alice Helliwell and Kevin Loi-Heng	Concepts and Classification Algorithms: A Case Study Involving a Large Language Model
Peter Tsu	The Ethical Frame Problem and Moral Perception Situated in a Form of Life
Rayane El Masri and Aaron Snoswell	Towards Attuned AI: Integrating Care Ethics in Large Language Model Development and Alignment
Claas Beger	Towards AI Collaborators: Exploring Goal, Value and Role-Based Alignment
Markus Pantsar	Artificial and human mathematical reasoning
Enrique Aramendia Muneta	AI and consciousness: How long is the shadow of epiphenomenalism?
Adrian Cussins	life and intelligence are different responses to the same cognitive challenge
Huseyin Kuyumcuoglu	Contractualist Solution for the Fairness Impossibility Theorem

AUTHORS	TITLE
Sabato Danzilli	The writing of the query as a hermeneutical act
Elina Nerantzi	Between persons and things: AI agents in Criminal Law
Jonathan Pengelly	Moral Cartography and Machine Ethics
Cecilia Vergani	The Social and Political dimension of Work: Technological Unemployment as a Threat to human cooperation, social integration and solidarity.
Konstantinos Voukydis	Phenomenal Consciousness in the Age of Large Language Models
Oliver Hoffmann	Framing Subjects and Objects
Frieder Bögner	Attention economy, exploitation and recognition-based harms
Qiantong Wu	The Philosophical Zombie and The Possibility of AI Consciousness in Large Language Models
Rokas Vaičiulis	The Externalist Implications of Machine Learning Epistemology: Empirical Knowledge and Its Social Dimension in the Accounts of C. Buckner and M. Pasquinelli
Hannah Louise Mulvihill, Taís Fernanda Blauth, Oskar Josef Gstrein and Andrej Zwitter	A systematic review of values integral to ethical design frameworks for the governance of artificial intelligence
Daniel Hromada	Prelude to Hermeneutics of Latent Spaces