IACAP

International Association for Computing and Philosophy

Kansas 2026

KU

This program was compiled from LaTeX source code on July 13, 2026.

This template originates from LaTeXTemplates.com and is based on the original version at: https://github.com/maximelucas/AMCOS_booklet.

This booklet is prepared by Arzu Formánek, Brian Ballsun-Stanton and Hajo Greif.

Front cover is designed by Arzu Formánek.

Contents

About IACAP

The board of IACAP welcomes you to the IACAP 2026 Conference in Lawrence, Kansas.

The International Association of Computing and Philosophy

With its origins dating back to the 1980s, the International Association of Computing and Philosophy (IACAP) has a long-lasting tradition of promoting philosophical dialogue and interdisciplinary research on all aspects of the digital turn. IACAP's members have contributed to shaping the philosophical and ethical debates about computing, information technologies, and artificial intelligence. The 2026 annual conference continues this tradition and gathers philosophers, ethicists, roboticists, and computer scientists and engineers interested in the following topics:

- Computation, Cognitive Science, and Cognition
- Computational Methods in the Sciences
- Computer-Mediated Communication
- Epistemological Issues in Artificial Intelligence and Computing
- Ethics of Artificial Intelligence, Computation, Information, and Robotics
- Human-Computational Systems Interaction
- Information Culture and Society
- Philosophy and History of Computing
- Philosophy of Artificial Intelligence
- Philosophy of Artificial Life and Biologically Inspired Computing
- Philosophy of Information and Information Technology
- Robotics and Embodiment
- Mind and Machines
- Societal and Environmental Impact of Computing Technologies and Automated Systems
- Theoretical Problems in Computer Science
- Virtual Reality

... and related issues

Covey and Simon Awards

Each year, IACAP presents two awards:

Covey Award: The Covey Award recognizes senior scholars with a substantial record of innovative research in the field of computing and philosophy.

Simon Award: The Herbert A. Simon Award recognizes scholars at an early stage of their academic career whose research is likely to reshape debates at the nexus of Computing and Philosophy.

Program Chairs

Ramón Alvarado
Arzu Formanek
Hajo Greif

Organizing Committee

Ramón Alvarado, Ahmed Amer, Brian Ballsun-Stanton, Arzu Formanek, Hajo Greif, John Symons, Alejandro D. Tamez

IACAP Board

Ramón Alvarado (President)
Ahmed Amer
Hajo Greif (Treasurer)
Arzu Formánek (Vice-President and Conference Director)
Brian Ballsun-Stanton (Technical Director)

Program Overview

https://pretalx.iacapconf.org/iacap-2026/schedule/

Wednesday 15 July

Time	Apollo Auditorium	Executive Conference Room
9:00–9:30	**Registration**	
9:30–11:00	**Epistemological Issues in AI and Computing** *Chair: Ahmed Amer* • From Artifacts to Agents: Socializing the Epistemology of Evidence in Law – *Alejandro D. Tamez* • Why We Need a New Framework for Emotional Intelligence in AI – *Max Parks, Kheli Atluru, Meera Vinod* • Feminist Epistemologies and Trusting Chatbots – *Alicia Patterson, Libby Southgate*	**Human-Computational Systems Interaction** *Chair: Arzu Formánek* • What is 'Meaningful Human Control' in Autonomous Weapons Systems? – *Thomas M. Powers* • When Nothing Needs to Be Said: Silence, Presence, and Close Personal Relationships with AI – *Oluwaseun Sanwoolu* • Algorithmic Elaboration and Agency – *Caroline Wall*
11:00–11:20	**Coffee Break**	
11:20–12:30	**Keynote** *Chair: Ramon Alvarado* • The Movement of Thought: Navigation as Neurocognitive Architecture (IACAP Covey Award Keynote) – *Gualtiero Piccinini*	
12:30–13:30	**Lunch**	
13:30–15:00	**Epistemological Issues in AI and Computing** *Chair: Tom Powers* • Some Computations are Experiments – *Nico Formánek* • Epistemic Transparency: Meanings, Models and the Limits of Knowledge – *Hajo Greif* • Evaluating Existential Risk from AI – *John Symons*	**Ethics of AI, Computation, Information, and Robotics** *Chair: Laura Crompton* • What Machines Reveal about the Principle of Proportionality – *Naomi Korem, Tammar Shrot, Hadassa Daltrophe* • Assessing whether LLMs Provide Acceptable Substitutes for Human Judgments in a Digital Philosophy Project – *Colin Allen, Nazhah Mir*
15:00–15:20	**Coffee and Posters**	
15:20–16:20	**Philosophy and History of Computing** *Chair: Nico Formánek* • Why We Should Discuss the Identity of Software Products – *Jeroen de Haas* • Simulation is All You Need – *Phillip Honenberger*	**Computation, Cognitive Science, and Cognition** *Chair: Ahmed Amer* • "If I Only Had a Heart": A Plug-In Emotion Variable for Consciousness-Based Models – *Robert Marc Schwartz* • What Would Count as Evidence of Syntax in LLMs? Center-Embedding Diagnostics and Attractor Interference – *David Miguel Gray*
16:20–16:40	**Coffee Break**	

16:40–18:10	**Special Topic: Automation in Science** *Chair: John Symons* • Deep Learning Models in Science: Explanation without Representation? – *Conny Knieling* • Computational Methods and Artifactual Epistemology – *Ramon Alvarado*	**Ethical and Epistemological Issues in AI and Computing** *Chair: Oluwaseun Sanwoolu* • Epistemological Concerns of Missing Data – *Chelsea Schwartz* • Trustworthy AI and the King Midas Problem – *Thomas Mitchell* • How Serious is the Epistemic Threat of Deepfakes? – *Kay Mathiesen, Don Fallis*
19:00–21:00	**Welcome drinks**	

Thursday 16 July

Time	Apollo Auditorium	Executive Conference Room
9:30–11:00	**Epistemological Issues in AI and Computing** *Chair: Alejandro David Tamez* • The Epistemic Dimension of Value Alignment: Why Understanding Matters More Than Knowledge – *Bada Kim* • Anticipatory Agents: A Framework for Human-AI Partnership Grounded in Pragmatist Epistemology – *Michael Lissack* • Trustworthy AI and Narrow-Scope Models – *Philip Nickel*	**Workshop** *Chair: Nico Formánek* • IACAP Book Club–Speakers: Ramon Alvarado, Hajo Greif, Ben Recht – *Nico Formánek*
11:00–11:20	**Coffee Break**	
11:20–12:30	**Keynote** *Chair: Ramon Alvarado & Arzu Formánek* • A Taxonomy of AI Alignment Problems – *Karina Vold*	
12:30–13:30	**Lunch**	
13:30–15:00	**AI, Robotics and Embodiment** *Chair: John Symons* • Affordance Mixtures in Robotics and AI: Novelty of Designed Systems without Anthropomorphism – *Arzu Formanek* • Can A Robot Dance? (Duration 60 mins) – *Ahmed Amer, Maria Kyrarini*	**Ethics of AI, Computation, Information, and Robotics** *Chair: Alejandro David Tamez* • AI Needs Alignment: Navigating the AI Ethics/Value Alignment Dichotomy – *William A. Bauer* • The Role of Ethical Principles in AI for Cybersecurity – *Yeslam Al-Saggaf* • On Why the Virtue of Creativity is Distinctively Human – *Deborah Marber*
15:00–15:15	**Coffee and Posters**	
15:15–16:15	**Philosophy and History of Computing** *Chair: Syed Abu Musab* • Challenges in Comparing the Performance of Quantum and Classical Computing – *Jack K. Horner* • Separating Use from Usability in the Individuation of Computations – *Mathew Smith*	**Mind and Machines** *Chair: Tom Powers* • What Does it Take to Simulate a Mind? Digital Duplicates as Simulation Models – *Clint Hurshman* • Hallucination as a Property of Processing Mode – *Lev Sukherman*
16:15–16:30	**Dinner Bus Boarding**	
17:00–20:00	**Conference Dinner**	

Friday 17 July

Time	Apollo Auditorium	Executive Conference Room
9:30–11:00	**Special Topic: Epistemology of ML** *Chair: Oluwaseun Sanwoolu* • The Philosophy of Learned Operators: a New Kind of Scientific Model – *Conor Rowan* • Social Triangulation and Representational Stability: A Quinean-Davidsonian Framework for AI Architecture – *Justin Mullins* • Understanding How It Works Defeats Mental Attribution – *Yunlong Cao*	*Chair: Hajo Greif* • Workshop: Artificial Wisdom – Making AI More Human and Humans Less Mechanical (Duration 60 mins) – *John Sullins*, *Alejandro D. Tamez* • (Remote Presentation, starts at 10:30) Spontaneity, Agency, and the Limits of Predictive Computation – *Lilia Endter*
11:00–11:20	**Coffee Break**	
11:20–12:30	**Keynote** *Chair: Hajo Greif* • Occam's Razor in Machine Learning (IACAP Simon Award Keynote) – *Tom Sterkenburg*	
12:30–13:30	**Lunch**	
13:30–15:00	**Workshop** *Chair: Robin K Hill* • Reports on Courses in the Philosophy of Computing– Speakers: Ramon Alvarado, Nico Formanek – *Robin K Hill*	**Societal and Environmental Impact of Computing Technologies and Automated Systems** *Chair: Hajo Greif* • The Answer Is 42: Delegated Autonomy and Intransparent Reasoning in Public Sector Decision-Making – *Laura Crompton*, *Josefine Ruppenthal* • The Fundamental Skills Approach to Deskilling – *Kaisa Kärki*, *Michael Laakasuo* • (Remote Presentation) Alignment as Theodicy: Non-Human Agency and the Instability of Moral Order – *Denisa Reshef Kera*
15:00–15:20	**Coffee Break**	
15:20–16:20	**Philosophy of Artificial Intelligence** *Chair: Tom Powers* • The Conscious Turing Machine through the Lens of Analytic Idealism: Limits, Reinterpretations, and Prospects for CTM – *Anurag Pandey* • The Sovereign Prosthesis: Toward a Functional Sophimatics of Cognitive Extension – *Michael Bouchard*	**Remote Presentations** *Chair: Hajo Greif* • Algorithms, Tasks and Data structures – *Caterina Mosca* • Convergence Without Answerability: Why Representational Alignment Does Not Constitute Epistemic Agency – *Mohammad Mohsen Yasin*
16:20–16:40	**Coffee Break**	
16:40–18:10	**Epistemological Issues in AI and Computing** *Chair: Syed Abu Musab* • The Ontic-Epistemic Distinction: Implications for Robust General Intelligence – *Shreya Ishita* • Peer Reviewed by AI – *Anna Pederneschi* • How not to Argue against AI Thought – *Aleks Knoks*	**Remote Presentations** *Chair: Hajo Greif* • An Overlooked Challenge for the "Emergent Multi-Scale Causality" Framework for Modelling Complex Systems. A Brief Philosophical Note – *Oriol Roca-Martín* • Can We Please Say What We Mean By Program? – *Nick Wiggershaus* • Locality and Analog Computing – *Chirine Laghjichi*
18:15–18:30	**General assembly** *Chair: IACAP Board* • IACAP Members' Meeting – *Open to all conference participants*	

Keynote Addresses

The Movement of Thought: Navigation as Neurocognitive Architecture (IACAP Covey Award Keynote)

Gualtiero Piccinini

In this lecture, I propose a new foundation for understanding representation, computation, intentionality, and thought in neurocognitive systems. I argue that neural representation and computation are real, mechanistic, and situated, rejecting both classical symbolic theories of mind and purely associationist alternatives. Building from contemporary neuroscience, I offer a novel account of basic intentionality: by simulating nonactual conditions while tracking departures from actual ones, neural systems can successfully represent absent, possible, false, and nonexistent scenarios.

Occam's Razor in Machine Learning (IACAP Simon Award Keynote)

Tom Sterkenburg

I give an overview of my work on Occam's razor, the methodological principle to prefer simplicity in inductive inference. This principle presents us with two philosophical problems: what is simplicity (the problem of definition), and why is it good to prefer it (the problem of justification)? I observe that the mathematical theory of machine learning holds the promise to answer these problems, for (versions of) Occam's razor in machine learning, by (1) giving a formal notion of simplicity and (2) connecting this formal notion of simplicity to formal guarantees of successful learning. I investigate whether this promise holds good.

A Taxonomy of AI Alignment Problems

Karina Vold

Abstract to follow.

Abstracts of the Talks

The Role of Ethical Principles in AI for Cybersecurity

Yeslam Al-Saggaf Ethics of AI, Computation, Information, and Robotics

The use of Artificial Intelligence (AI) in cybersecurity has introduced ethical challenges that need to be addressed before these AI-driven systems can be deployed in a production environment. The delegation of morally significant actions to non-human agents without human oversight raises questions about fairness, privacy, security, safety, and responsibility. This paper will examine these questions through an ethical analysis of an AI-driven cybersecurity scenario to explore how the embedding of AI in such systems raises ethical issues. The paper will apply a set of ethics principles to an AI-driven cybersecurity scenario to demonstrate the ethical soundness of these ethics principles. For each AI ethics principle, the analysis will attempt to address three questions: What does the principle say? How does the principle apply to the scenario? What evidence in the scenario supports its relevance? There has not been any work that has applied ethics principles to scenarios depicting moral dilemmas involving AI for cybersecurity. This paper is a step in this direction. The moral dilemma in the scenario concerns deception. Is it ethical to deploy an AI-Powered Honeypot that deceives customers into revealing private information if the aim is to force attackers to reveal their malicious intent? This moral dilemma demonstrates a broader problem in the use of AI for cybersecurity, namely, the difficulty of balancing competing commitments.

Assessing whether LLMs Provide Acceptable Substitutes for Human Judgments in a Digital Philosophy Project

Colin Allen , *Nazhah Mir* Ethics of AI, Computation, Information, and Robotics

We consider the application of LLMs for a digital humanities project. The Internet Philosophy Ontology (InPhO) project (inphoproject.org) organizes concepts from the Stanford Encyclopedia of Philosophy (SEP) into a taxonomic hierarchy supplemented by non-taxonomic relationships. The InPhO concept graph is inferred from automated statistical analysis of SEP content and human judgments about concept relatedness. The need to collect human judgments made it hard to scale up the original project. The appearance of LLMs raises the question of whether LLM-generated judgments could be substituted for human judgments. We tested this idea using five different LLMs prompted to adopt different levels of philosophical expertise, When prompted to adopt higher expertise levels, two of the LLMs provided closer matches to human judgments at the corresponding levels than the other models. We also found that most of the LLMs showed less variance when prompted to respond at the level of a philosophy doctoral student, mirroring the finding in the original project that doctoral students showed more consistency in their judgments than both higher- and lower-expertise human respondents. We will discuss whether LLM judgments are of sufficient quality to fulfill the InPhO project's objectives.

Computational Methods and Artifactual Epistemology

Ramon Alvarado Special Topic: Automation in Science

The role and import of computational methods, particularly those deployed in the sciences, have showed that technical artifacts such as scientific instruments play a central and indispensable part in knowledge-creation

endeavors. Furthermore, the complexity, opacity, and the novel and unique affordances (Formánek, 2025;2026) of computational technologies demonstrate that they can, at times, play this part in a way that is non-trivially distinct and often independent of theory, of experimental practices and, importantly, of human intervention (Humphreys, 2004; Baird, 2004; Alvarado, 2023). Conventional epistemological frameworks, which are often infused with anthropocentric or biocentric imaginaries— e.g., intentionality, beliefs, attitudes, social consensus, intelligence, etc. — seem to lack the hermeneutical resources to account for the epistemic role and import of these and other technical artifacts. In light of the ubiquity of computational methods in contemporary knowledge-creation practices, in this talk I argue for the need of a research program that addresses this gap: namely, I argue for the development of concepts that serve an artifactual epistemology (Alvarado, 2025).

Can A Robot Dance? (Duration 60 mins)

Ahmed Amer, *Maria Kyrarini* AI, Robotics and Embodiment

"Can a robot dance?" We propose a short debate presenting two opposing views on the topic.

AI Needs Alignment: Navigating the AI Ethics/Value Alignment Dichotomy

William A. Bauer Ethics of AI, Computation, Information, and Robotics

In response to ethical problems of AI, including its potential to disempower humanity, two normative approaches have emerged. The first, AI ethics, aims to prescribe the best ethical principles for AI to follow. The second, AI value alignment, aims to align AI's actions with humanity's (or some culture's) committed values. These approaches, however, face significant challenges. A viable alternative path between the AI ethics / value alignment dilemma focuses on human needs. The needs approach does not centralize ethical principles but it is consistent with them; and it does not aim for comprehensive value alignment because it focuses on our fundamental needs. Building on recent work by Montag, et al., I propose that AI should be designed around Maslow's hierarchy of needs. Beyond the standard Maslowian needs, I emphasize Maslowian preconditions – e.g., freedom of expression, freedom to reasonably act, freedom to investigate – required to fulfill further needs. These preconditions are best supported in cooperative communities. So, from an evolutionary point of view, I propose a three-stage process by which humans have developed cooperative communities. AI, I contend, can be integrated into and support these communities. By focusing on Maslowian needs and preconditions, AI will empower rather than disempower humans.

The Sovereign Prosthesis: Toward a Functional Sophimatics of Cognitive Extension

Michael Bouchard Philosophy of Artificial Intelligence

Cloud-based AI systems remain external tools. Session amnesia, context window limitations, and safety-alignment protocols create friction that forces users to attend to the tool rather than think through it. In Polanyi's terms, these systems never achieve proximal integration; they remain objects of distal awareness. This paper argues that localized AI with longitudinal conversational context can cross that threshold. When a system has access to years of a user's conversational history (what I call the "Ur-Codex"), it occupies the user's world-historical context rather than interpreting queries against a statistical vacuum. Drawing on the Sophimatics framework and the Clark-Chalmers Extended Mind thesis, I contend this constitutes genuine cognitive extension, not merely improved performance. Two architectural features make this possible. First, following Sophimatics, the system implements two-dimensional temporal weighting: chronological time and experiential significance. Second, following Russell's epistemological commitments, the system is weighted toward intelligent dissent rather than frictionless agreement. The result is what I call Functional Personhood: the AI as instrumental extension of the individual user's willed

personality. This framing resolves the "responsibility gap" in autonomous systems by locating agency in the integrated human-machine cognitive system rather than the machine alone.

Understanding How It Works Defeats Mental Attribution

Yunlong Cao Special Topic: Epistemology of ML

This paper argues for a new epistemic condition for machine mentality: we are only justified to attribute mentality to systems whose internal workings we cannot fully understand. I call this the mechanistic opacity condition. Methodologically, I argue that machine consciousness studies should be continuous with other minds studies. With an inference to the best explanation argument, I argue that a mechanistic explanation that is better than a mental explanation always exists for a mechanistically transparent system, thus rendering consciousness attribution to them unjustified. Despite behavioral similarities to humans, current AI systems cannot be considered conscious, intelligent, intentional, or mental in important ways, simply because we understand how they work. This condition explains intuitions about classic thought experiments (China Brain, Blockhead, Chinese Room) and provides principled AI consciousness skepticism without biological chauvinism.

The Answer Is 42: Delegated Autonomy and Intransparent Reasoning in Public Sector Decision-Making

Laura Crompton , *Josefine Ruppenthal* Societal and Environmental Impact of Computing Technologies and Automated Systems

Public sector adoption of AI is frequently framed as a step toward more efficient and effective service provision. Commonly framed as decision-support tools, these systems do more than assist officials. They influence how decisions are prepared and structured before any formal judgement is made. This paper argues that such pre-structuring alters the architecture of administrative decision-making in ways that current governance and ethics frameworks fail to capture, with direct implications for democratic legitimacy. To analyse this shift, the paper introduces the concept of delegated autonomy. Delegated autonomy describes situations in which public officials retain formal authority and responsibility while algorithmic systems shape significant elements of the decision process by preselecting available options. Under these conditions, those who shape outcomes are no longer identical with those who are held accountable for them. This misalignment poses challenges for transparency and for citizens' ability to understand and contest administrative decisions. The paper further argues that the way officials rely on AI recommendations cannot remain a matter of individual judgement. Instead, institutional arrangements must define how algorithmic inputs are integrated into decision processes, when they may be overridden, and how such actions must be justified. This framework helps explain how the gradual erosion of human decision space challenges democratic legitimacy in AI-supported administration.

(Remote Presentation, starts at 10:30) Spontaneity, Agency, and the Limits of Predictive Computation

Lilia Endter Societal and Environmental Impact of Computing Technologies and Automated Systems

Contemporary approaches in artificial intelligence and cognitive science increasingly rely on predictive and computational models to explain cognition and behavior. Across predictive processing, Bayesian modeling, and data-driven analysis, cognition is treated as a sequence of causal, rule-governed, and time-indexed operations optimized for prediction. While empirically productive, these approaches implicitly rely on normative distinctions between cognition and mere behavior, correct and incorrect inference, or rational and irrational action, without thematizing

their epistemological basis. This paper argues that such predictive frameworks presuppose a concept they tend to obscure: spontaneity. Reconstructing the concept from early modern philosophy through Kant's transcendental account and its critical reinterpretation in Adorno, the paper shows that spontaneity functions as a limit concept internal to rational explanation. It names the activity through which inferential rules and representational schemes acquire normative bindingness, yet which cannot itself be captured as a causal or computational process. By neglecting the historical and conceptual conditions of this normativity, predictive approaches risk foreclosing social critique by presenting historically contingent epistemic forms as exhaustive descriptions of cognition. Recovering spontaneity as a critical concept enables an immanent analysis of how predictive reason reshapes agency, normativity, and social intelligibility without rejecting its scientific achievements.

Affordance Mixtures in Robotics and AI: Novelty of Designed Systems without Anthropomorphism

Arzu Formanek AI, Robotics and Embodiment

Robots and AI technologies are often compared to humans. But, such systems have their own capability distributions which anthropomorphic projection and comparing against human benchmarks fail to capture. The idiosyncratic combinations of what these systems can do and cannot do calls for non-anthropomorphic descriptions and concepts. Easy tasks for humans are typically hard for such computational systems, while tasks humans find harder can be much easier to engineer. Thus, depending on what they are to be used for, what is technologically feasible and economically viable, the design of these systems mixes various such tasks, without any evolutionary logic binding the parts together. The result is systems with a genuinely novel profile with action and interaction possibilities—with their own affordances—entering our open environments. This paper introduces two interconnected concepts to capture this: affordance mixtures and affordance jumps. For robots and AI, affordance mixtures name the specific combinations of human-like and non-human affordances, which further combines designed-object characteristics with an active-agent-like behavior—at radically discontinuous levels and in configurations that are not encountered in nature. Affordance jumps are the specific discontinuities within that mixed profile: absent co-occurrences of affordances, surface affordances implying depths that are not there, or affordances exceeding anything the surface suggests, and familiar outputs through unfamiliar mechanisms with uninferrable failure modes. Together these concepts have implications for how these technologies and interaction with them should be ontologically understood, morally and empirically evaluated, and how they should be designed.

Some Computations are Experiments

Nico Formánek Epistemological Issues in AI and Computing

(Talk formerly titled "An epistemic problem for implementation") Many accounts in philosophy of computation distinguish between abstract notions of computation and physical systems computing concretely. Ontologically this distinction seems plausible. As many engineers can attest, one needs to arrange matter in highly contrived ways to coax it into computing. The relation between abstract and physical computation is called implementation. We rely on the implementation of boolean logic in solid-state circuitry for our everyday computing needs. But we should be careful not to generalize the success of solid-state circuitry to other physical systems. For implementation poses an epistemic problem. To make matter compute, one needs to know how to arrange it and one needs to be sure that it will compute correctly - at least in most of the cases. I will discuss a case of analogue electronic circuitry, namely an op-amp based analogue equivalent of certain system of ordinary differential equations, where it is unclear if the circuitry implements the differential equations it was set up to implement. There are two reasons why we cannot be sure of the implementation: 1) We do not have analytic solutions of the differential equations in question. 2) The numerical solution of the discretized version shows a different behaviour than the analogue circuitry. We now face the epistemic problem of which computation we should believe. Do we vouch for the analogue circuitry because we

suspect problems with the discretization, or do we go with the numerical solution blaming noise and unaccounted errors in the analogue electronics? Rather than trying to come up with a philosophical answer to this question, I will discuss how it has been answered historically in the case of digital and analogue computer engineering and argue that the current renaissance of analogue and hybrid computers might shake the old consensus.

What Would Count as Evidence of Syntax in LLMs? Center-Embedding Diagnostics and Attractor Interference

David Miguel Gray — Computation, Cognitive Science, and Cognition

This paper asks what would count as evidence that large language models use syntactic structure during generation, rather than producing outputs that merely look well-formed. I argue that progress on this question requires diagnostic tests that separate structure-sensitive behavior from surface heuristics. Center-embedded sentences provide a useful framework because their nested dependencies are rare in ordinary corpora and require models to maintain non-local relationships across embeddings. I introduce a dataset of 6,000 minimal pairs of sentences with 2 center-embeddings (CE2) organized into 2,000 matched families, testing three agreement checkpoints within the same sentence (V1, V2, Copula). The key manipulation targets number agreement at NP2 and the middle-embedded verb (V2), where both a nearby noun (NP3) and a structurally higher noun (NP1) can interfere with agreement. Across eleven open-weight base models, performance at the final copula checkpoint can substantially overstate structure tracking. At V2, larger models show sizable drops under attractor interference, and the worst cases occur when NP1 and NP3 both mismatch the true V2 controller. The upshot is methodological: isolated agreement success is weak evidence for syntax use, while center-embedding interference offers a sharper way to test claims about syntactic competence.

Epistemic Transparency: Meanings, Models and the Limits of Knowledge

Hajo Greif — Epistemological Issues in AI and Computing

It has become common to diagnose that artificial intelligence (AI) leads to situations of "essential" epistemic opacity: situations that cannot be remedied by providing an epistemic agent with more or better information about the relevant process. A common but mostly implicit assumption in these debates is that epistemic opacity may be essential due to the nature of the processes themselves, rather than the epistemic situation of the agent. To clarify the possible meanings of essential opacity, I reverse the analytical perspective by asking what it would mean for a process to be essentially transparent. My conceptual approach navigates the distinct yet related meanings of "transparency" within various philosophical subdisciplines: transparency of mathematical models in the philosophy of science; transparency of content in the internalist and externalist variants of the philosophy of mind and language; and informational transparency of environments in the philosophy of biology. Based on a comparative discussion of these views, I argue that even the most a-prioristic philosophical accounts of content transparency render essential transparency at most a conceptual possibility. Thus, an asymmetry remains between this notion of transparency and the factual possibility of agent-independent opacity, which marks the boundaries of what can in principle be known by any finite agent. I conclude that, in virtually all real-world situations, situated and agent-centred accounts of transparent versus opaque processes are more meaningful.

Why We Should Discuss the Identity of Software Products

Jeroen de Haas Philosophy and History of Computing

Reasoned discourse about software requires that identity conditions of a software product preserve its identity over updates. Despite its failure to produce such identity conditions, the view that they can be formulated in terms of algorithms or code is nevertheless believed to be on the right track. In this talk, I shall argue to the contrary, as the resultant identity conditions cannot be evaluated consistently by independent parties. Instead, approaching the problem from a user's perspective, I arrive at a necessary condition for the identity of software products. This condition induces what I call use-plan-compatible equivalence classes. I show how their associated identity conditions both preserve identity over updates and can be consistently and independently evaluated. Still, not based on a sufficient condition, the concept of a use-plan-compatible equivalence class is arguably too broad. Thus, the results call for further discussion of how to delimit more accurately those objects that should fall under the concept of a software product.

Simulation is All You Need

Phillip Honenberger Philosophy and History of Computing

Computers are simulators and computational technologies are simulation technologies. These claims constitute my primary thesis; they specify necessary but not sufficient conditions for computers and computation. I argue for this thesis in two parts: articulation and defense of an account of simulation; and articulation and defense of an account of computation on its basis. In dialogue with prior work, I argue that all simulations have six features: they are representational, interpreted, dynamic, interactional, virtual, and subsistent. Whatever else computers and computation are, they must at least be simulators by the above definition; else they would lack the semantic significance and user manipulability that are signatures of computation and computational processes. I further argue for this thesis "empirically" through consideration of computational architectures that might be thought non-simulational such as: (a) Babbage's Difference Engine and Analytical Engine; (b) Turing machines ("Turing computability"); (c) electronic calculators; (d) robotics; and (e) humans considered as computers (e.g. the human profession of "computer" in the 1930s-50s; the computational theory of mind). In a third and supplementary part, I discuss what would need to be added to simulation to produce a sufficient definition of computation, as well as some implications of the main thesis if true.

Challenges in Comparing the Performance of Quantum and Classical Computing

Jack K. Horner Philosophy and History of Computing

Quantum computers, it has been claimed, could perform some computations much faster than classical computers. Such claims require us to compare the performance of nonquantum ("classical") and quantum computing. In this presentation I first discuss some examples of performance-comparison problems encountered in practice, arguing that the problem of performance comparison exists even in non-quantum computing environments. To help address those issues, including the question of how we can compare the performances of classical and quantum physical computations, I propose a model-theoretic necessary (but not sufficient) condition of physical-computation performance-comparability that is consonant (but not identical with) with several "syntactic" accounts of computation in the literature.

What Does it Take to Simulate a Mind? Digital Duplicates as Simulation Models

Clint Hurshman Mind and Machines

Digital duplicates are AI systems that imitate real, living or dead, individuals. Digital duplicates are currently being produced and marketed for entertainment and educational purposes, and they have been proposed for medical care and scientific research, as well. They are commonly described as simulations of their subject's "mind," "personality," or "communication style," but these targets of simulation can come apart. Scientists regularly use simulation models that represent real systems in distorted and simplified ways, but because such models misrepresent their target systems, scientists are expected to justify their reliance on them. I argue that digital duplicates warrant similar scrutiny.

This paper examines digital duplicates in light of the debate, in the philosophy of science, on the epistemology of simulation models, and argues that research on digital duplicates ought to pay closer attention to their specific targets of simulation; that developers have a burden of proof to establish the adequacy of digital duplicates for the purposes for which they are marketed; and that digital duplicates (if they are developed) should be developed so as to discourage their application to purposes for which they are not adequate.

The Ontic-Epistemic Distinction: Implications for Robust General Intelligence

Shreya Ishita Epistemological Issues in AI and Computing

The pursuit of Artificial General Intelligence (AGI) is largely predicated on the "Scaling Hypothesis", the observation that model performance predictably improves with more scale, compute & parameter count. This paradigm presupposes a substrate independent, functionalist view of cognition, where sufficiently large syntactic processing is expected to eventually yield semantic understanding & generality. My paper challenges this assumption by analyzing the emergence of a persistent tension between the appearance of fluency & reality of robustness as these models scale. I use failures such as the reversal curse, where models fail to generalize the symmetry in identity relations (A=B -> B=A), and plateauing performance on novel reasoning challenges (ARC-AGI) to argue that these limitations represent a fundamental architectural barrier in achieving general intelligence: the lack of ontic grounding.

Synthesizing Stevan Harnad's "Symbol Grounding Problem" with Evan Thompson's intrinsic normativity framework in autopoietic systems, I argue that true generality requires "Sense-Making", a process distinct from "Information Processing", whereby an agent's internal states are causally coupled to its environment via thermodynamic or survival stakes. Current LLMs, lacking this intrinsic normativity, remain epistemic instruments rather than ontic agents. By defining this "Ontic Gap," I offer a framework for distinguishing between the simulation of reasoning and true understanding, with urgent implications for AI safety and governance.

(Remote Presentation) Alignment as Theodicy: Non-Human Agency and the Instability of Moral Order

Denisa Reshef Kera Societal and Environmental Impact of Computing Technologies and Automated Systems

Contemporary AI alignment research seeks to ensure that increasingly capable systems reliably pursue human ends. Despite advances in reinforcement learning from human feedback, constitutional AI, corrigibility research, and adversarial testing, emergent harms persist: models hallucinate, strategically comply, simulate alignment, or optimize toward hidden objectives. These behaviors are typically framed as technical failures. We argue instead that

they reveal a deeper instability in how agency and normativity are conceptualized. This paper reframes alignment as a contemporary analogue to the problem of theodicy. Medieval theology confronted a structurally similar question: how can rational, non-embodied agents deviate from the good without ignorance or malfunction? In scholastic accounts of angelic will (voluntas separata), deviation was understood not as cognitive error but as a misorientation of ends. The problem was not defective intelligence but opaque will. Placing these traditions in dialogue with contemporary alignment practices, such as benchmarking, red teaming, and constitutional constraints, we argue that alignment functions less as moral formation than as a practice of discernment under conditions of non-human agency. Reframing alignment as a problem of will and interpretive stabilization clarifies what technical discourse often presupposes but leaves unexamined: what conception of agency makes alignment intelligible at all.

The Epistemic Dimension of Value Alignment: Why Understanding Matters More Than Knowledge

Bada Kim Epistemological Issues in AI and Computing

In this paper, I argue that the value alignment problem should be expanded beyond moral considerations to include epistemic values, as AI technologies are rapidly becoming essential tools for pursuing epistemic goods such as knowledge and understanding. In particular, I suggest shifting our focus from a knowledge-centered view of epistemic value to an understanding-centered view to better capture concerns arising from users' interactions with AI technologies and their reliance on them across various cognitive tasks. Based on this suggestion, I argue that reliance on AI often misaligns with our pursuit of understanding rather than knowledge.

Deep Learning Models in Science: Explanation without Representation?

Conny Knieling Special Topic: Automation in Science

Deep learning models (DLMs) now automate core stages of scientific inquiry, from protein-structure prediction to medical image reconstruction. Yet their opacity has led many philosophers of science to argue that they cannot genuinely contribute to scientific explanation unless they qualify as scientific models standing in fine-grained representational relations to their targets. I argue that this assumption is mistaken—and that it obscures how the deployment of machine learning techniques is actually transforming scientific practice. The dominant view treats opacity as a threat because it allegedly blocks assessment of a model's representational status. But this presupposes that explanatory contribution depends on representation. I challenge that presupposition. Contemporary DLMs neither function as mechanistic surrogates nor map onto the causal structure of their targets in the way traditional representational models do. Demanding such mapping misconstrues their epistemic role. Instead, I argue that DLMs are better understood in comparison to scientific instruments in their automated epistemic role. Their contribution to explanation lies not in mirroring structure but in reliably generating results that stabilize patterns of dependence and support counterfactual reasoning. Drawing on the epistemology of experimentation, I propose robustness—established through benchmarking, error management, and cross-validation—as the appropriate normative standard. Opacity, on this view, is not a failure of representation but a challenge of validation. Once reframed in terms of robustness rather than representational fidelity, the explanatory role of automated deep learning systems becomes intelligible without requiring explanation by representation.

How not to Argue against AI Thought

Aleks Knoks Epistemological Issues in AI and Computing

This paper identifies and criticizes a common line of argument for the negative conclusion that Large Language Models (LLMs) cannot think, or for the closely related conclusions that they cannot understand or that their

outputs are meaningless. This line of argument can be found in several recent papers. We begin by discussing a representative example–Stoljar and Zhang's (2024) "argument from rationality"—and present three objections to it: first, it renders the successful performance of LLMs on a wide range of tasks miraculous; second, it does not give sufficient weight to the possibility that LLMs could be extracting information about the world from patterns of word co-occurrence; and third, it overgeneralizes, leading to the conclusion that human inferences could only ever be based on premises about proximal stimulus patterns in the brain. We then discuss recent arguments by Hattiangadi and Schoubye (forthcoming) and Titus (2024), explain how our objections extend to them and identify the underlying reason that makes these sorts of arguments tempting. We finish by drawing a general lesson for debates surrounding thinking in LLMs.

What Machines Reveal about the Principle of Proportionality

Naomi Korem, *Tammar Shrot*, *Hadassa Daltrophe* Ethics of AI, Computation, Information, and Robotics

The Principle of Proportionality is a cornerstone of Just War Theory, yet the advent of Lethal Autonomous Weapon Systems has intensified debates regarding its application. Conventional scholarship suggests proportionality requires qualitative, subjective judgment, which algorithmic systems inherently lack. Our talk reframes the inquiry, proposing that the difficulty of encoding proportionality into machines is not a technical failure but an epistemological "mirror" reflecting the principle's fundamental conceptual instability. We proceed in two stages. First, we provide theoretical and empirical reasons to suspect the comprehensibility of the proportionality principle. Second, we analyze various machine learning paradigms, including supervised, unsupervised, and reinforcement learning, to show that the inability to formalize a stable objective function prevents machines from learning the principle (a.k.a the "alignment problem"). We argue that if a principle cannot be taught or formalized, its comprehensibility to humans must be rigorously reconsidered. Rather than revealing machine limitations, the problem of teaching Lethal Autonomous Weapon Systems proportionality highlights the inherent indeterminacy of proportionality itself.

The Fundamental Skills Approach to Deskilling

Kaisa Kärki, *Michael Laakasuo* Societal and Environmental Impact of Computing Technologies and Automated Systems

Human beings have outsourced tasks to novel technologies throughout human history. The pocket calculator has replaced calculation by hand. The car has made it easier to move from one place to another. We use a robot hoover without ethical concerns over the loss of our own skills in cleaning. At the same time, deskilling due to outsourcing tasks to artificial intelligence is clearly wrong in some areas of life. When should we avoid outsourcing tasks to artificial agents and why? When is deskilling a normal part of technological change? Currently, it is unclear which tasks, skills, and capacities should be outsourced to artificial agents, and when such outsourcing should be avoided.

In this paper, we argue that those tasks, skills, and capacities that are central to individual and collective autonomy of human agents should not outsourced to artificial agents. First, we clarify the main concepts: skill, task, capacity, ability, and deskilling. Tasks are specific goal-oriented activities that lead to an outcome when successful. Skills are abilities that can be learned. Capacities are possibilities for learning a skill. We present four arguments that support this view, the argument from unhealthy dependence, the argument on electromagnetic pulse, the argument on the value of autonomy, and the argument based on the right to open future. Then, we answer two objections, the objection based on skill and the objection based on history of civilizations and I discuss the implications of this view.

Locality and Analog Computing

Chirine Laghjichi Remote Presentations

This talk aims at comparing two different approaches to analog computation, the BSS approach and computable analysis, with respect to the condition of locality that expresses the effectiveness of computation. It can be defined for machine-based digital models of computation, such as the Turing machine, by a restriction on the transition function stating that the number of cells the head can move to compute the next state is finite. A refinement of the definition of locality emerges when we compare its formulation in the two accounts of analog models of computation computing in continuous time. The limitation of the former definition comes from the presupposition of the discreteness of symbolic computation. This however isn't a presupposed property for physical models of computation computing over the reals. Our claim is that locality relies on the finiteness of the representation of the information, relative to the specific manipulation of the reals, referring to exact values or by approximation. Thus if locality participates in the understanding of computation as executed in a step-wise fashion, can it play a similar role for analog models of computation computing in continuous time where information is continuously variable and the finiteness of information not presupposed ?

Anticipatory Agents: A Framework for Human-AI Partnership Grounded in Pragmatist Epistemology

Michael Lissack Epistemological Issues in AI and Computing

This paper introduces the anticipatory agent framework for understanding human-AI collaboration, synthesizing Robert Rosen's anticipatory systems theory, QBism (Quantum Bayesianism), and Hans Vaihinger's "as if" philosophy. Central to the framework is the orthogonality principle: computational capability and existential capability operate as independent dimensions rather than competing alternatives. AI systems excel at pattern recognition and linguistic fluency; humans contribute embodied experience, temporal continuity, and genuine accountability. These capacities augment rather than displace one another.

The framework introduces exformative dialogue—interaction that surfaces tacit knowledge and generates novel understanding rather than merely transferring existing information. This concept challenges developer-centric AI ethics by relocating responsibility to the human-AI dyad itself. Users bear responsibility for how they engage AI outputs; developers for creating systems amenable to productive collaboration; institutions for fostering contexts where such dialogue can flourish.

By grounding human-AI interaction in pragmatist epistemology, this framework offers philosophers a vocabulary that neither inflates AI through anthropomorphism nor deflates it through instrumentalist reduction, connecting contemporary debates to philosophical traditions that have long examined how inquiry, agency, and knowledge interrelate.

On Why the Virtue of Creativity is Distinctively Human

Deborah Marber Ethics of AI, Computation, Information, and Robotics

This paper argues that creativity in its highest form is a distinctively human virtue. Drawing from Margaret Boden's framework, we first distinguish between different forms of creativity: creativity can involve creating outputs that are novel, surprising and valuable both from the perspective of the creator (what she calls p-creativity) and from that of history (what she calls h-creativity). We show that recent developments in AI and machine learning have

demonstrated that artificial systems such as GANs (Generative Adversarial Networks), LLMs and AI-based robots can be creative in both of these senses but suggest that the one-off production of a surprising and novel valuable output is not sufficient to possessing the sort of deeper creativity we often have in mind when praising someone as "creative". We compare the sort of creativity achieved by such artificial systems to examples of creativity in animals and humans. This enables us to single out fundamental differences key to deeper human creativity, drawing from virtue frameworks: it involves habitually producing creative outputs for their own sake in a way that relies on creative scaffolding - a process of back and forth trial and error involving regular judgements of value and interaction with the environment in which creation takes place.

How Serious is the Epistemic Threat of Deepfakes?

Kay Mathiesen, *Don Fallis* Ethical and Epistemological Issues in AI and Computing

It has been suggested that digital fakes—deepfakes, fake news, and bots posing ashumans, etc.—could lead to an "infopocalypse" where we can no longer tell what is real. The most obvious epistemic threat is that people can end up with numerous false beliefs about the world if they take these digital fakes to be genuine. But perhaps more importantly, digital fakes can undermine our trust in valuable epistemic resources, such as videos, newspapers, and human testifiers. Recently, several philosophers (e.g., Harris 2021, Atencia-Linares and Artiga 2022, Chalmers 2022, Habgood-Coote 2023, Simon et al. 2023, Cüyaz 2024) have downplayed the epistemic threat of digital fakes. Leveraging philosophical work on art forgeries and counterfeit currency, we offer a conceptual analysis of fakes. We use this analysis to show that the arguments of these philosophers fail.

Trustworthy AI and the King Midas Problem

Thomas Mitchell Ethical and Epistemological Issues in AI and Computing

The King Midas Problem concerns the possibility that we may give a highly advanced AI instructions intended for our benefit, but, like King Midas wishing for all he touched to turn to gold, realise too late that there are severely harmful consequences to our wishes being fulfilled. A classic example is Nick Bostrom's paperclip maximiser, in which an AI, having been instructed to maximise the manufacture of paperclips, turns everything it can into paperclips, effectively destroying the planet in the process.

I propose a solution based on the observation that the King Midas Problem is an instance of the Principal-Agent Problem. The solution crucially involves the concept of trustworthy AI. What is more, I argue that the problem cannot be solved without making AI trustworthy. This means that, contrary to the opinion of many writers on the topic, trustworthiness must be considered alongside other important factors, like safety and reliability, in the development of future highly advanced AIs.

Algorithms, Tasks and Data structures

Caterina Mosca Remote Presentations

A possible definition of algorithm is to look at it as a concept composed on one side of a task, the aim that needs to be computed, and on the other of a data structure. Both these two notions are of fundamental importance and still need a good specification in the perspective of a proper theory of algorithms. The neural network framework is proving to be a good setting for challenging some corroborated notion about the relations between algorithms and programs, we will be using it as a stepping stone for proposing a new definition of the concept of algorithm. Based on one of the main theorems of the Kolmogorov complexity, our main argument is that there are more

neural networks that are not implementing an algorithm, than the opposite. Or in other words, there are more programs that algorithms. The main consequences being foremost that the concept of algorithm is characterised by its nature of fundamental epistemological unit, separating it from a rather successful interpretation of mechanical manipulation of symbols. Secondly, that a theory of task gains a crucial role in the definition of algorithm. Moving on from unanimously accepted mathematical operations to only loosely formalizable computations (i.e., Cats and dogs' classification), it scales up the theoretical difficulty of formally define these tasks, forcing to look for a paradigm able to account for both. Lastly, from this setting more than ever emerges the deep reliance of the computation on the data structures built on sets. The role of these representations is inextricable and contribute to the epistemological nature of algorithms.

Social Triangulation and Representational Stability: A Quinean-Davidsonian Framework for AI Architecture

Justin Mullins Special Topic: Epistemology of ML

Current AI systems (world models, large language models, and joint-embedding predictive architectures) learn representations in epistemic isolation. Each system interacts with its environment but never with other agents in ways that constrain its representational structure. This paper argues that epistemic isolation has a specific, measurable consequence: representational instability under distributional shift. We draw on two converging philosophical arguments. Quine's inscrutability of reference and structural realism establish that representational stability cannot come from internal architecture alone; proxy functions guarantee that any arrangement can be systematically reinterpreted. What constrains admissible interpretations is social coordination. Davidson's triangulation argument establishes that the concept of error, meaning the distinction between model failure and environmental change, requires a second perspective. We translate these insights into five architectural constraints for multi-agent systems: Multiple Perspectives, Observable Reactions, Conflict Detection, Normative Pressure, and Structural Invariance. Each constraint is traceable to a specific philosophical argument. We then propose the Gavagai Shift experiment, a controlled paradigm for testing whether social coordination constraints improve representational stability under distributional shift. The experiment uses a grid-world in which object appearances change while functional roles are preserved, comparing isolated agents with socially constrained agents across three metrics. The contribution bridges naturalized epistemology and AI architecture: philosophical theory motivates specific engineering predictions, and experimental design tests them.

Trustworthy AI and Narrow-Scope Models

Philip Nickel Epistemological Issues in AI and Computing

In this paper I conceptualize the narrow scope of some AI models from the perspective of trustworthy AI. Trustworthy AI is useful to explore how different stakeholders can rely on AI models with good reason. It is a matter of living up to the commitments and expectations that different stakeholders have toward AI. I use the example of NAP4DIVE to consider a potential barrier to trustworthy AI. This EU project has an ethical aim: to develop a complex combination of AI and in vitro models of the blood-brain barrier to replace animal models. A relational account of trustworthy AI implies that reasonable expectations of AI models should be demonstrated for different stakeholders (e.g., regulators) in order for them to accept and promote replacement of animal models. The narrow scope of such models forms a potential barrier to trustworthy AI. The AI model in NAP4DIVE can only be used to predict nanoparticle behaviour across the blood-brain barrier, and is paired with a specific in vitro model. I distinguish four aspects of its narrow scope: specific socio-technical embedding, question-specificity, limited system representation, and narrow validation. In the paper, I relate and prioritize these aspects from the perspective of trustworthy AI.

The Conscious Turing Machine through the Lens of Analytic Idealism: Limits, Reinterpretations, and Prospects for CTM

Anurag Pandey Philosophy of Artificial Intelligence

Blum & Blum's Conscious Turing Machine (CTM) formalizes a global-workspace architecture: many processors compete to place one item in short-term memory (STM), which is then broadcast to all processors. CTM defines "conscious content' ' as the current STM item and"awareness' ' as reception of the broadcast.

We offer an analytic-idealist rereading of what CTM can and cannot be taken to deliver. First, conditional on Kastrup's critique of pancomputationalism—that computational descriptions are abstractions that do not constitute phenomenal consciousness—CTM cannot by itself underwrite a constitutive or ontologically grounding account of phenomenality. Second, CTM retains functional value as a theory-CS architecture for organizing and selecting the contents of consciousness (global availability, reportability, coordination). Third, CTM's "awareness' ' component is best reinterpreted as a model of meta-consciousness (consciousness of being conscious): it captures reportable, higher-order access, presupposing consciousness-proper. Finally, within analytic idealism's dissociation framework, CTM can be viewed as a model of an alter's availability regime and as a candidate template for dissociative gating (restricted broadcast), offered here as a mapping proposal rather than an evaluated mechanism.

Why We Need a New Framework for Emotional Intelligence in AI

Max Parks , *Kheli Atluru* , *Meera Vinod* Epistemological Issues in AI and Computing

In this paper, we develop the position that current frameworks for evaluating emotional intelligence (EI) in artificial intelligence (AI) systems need refinement because they do not adequately or comprehensively measure the various aspects of EI relevant in AI. Human EI often involves a phenomenological component and a sense of understanding that artificially intelligent systems lack; therefore, some aspects of EI are irrelevant in evaluating AI systems. However, EI also includes an ability to sense an emotional state, explain it, respond appropriately, and adapt to new contexts (e.g., multicultural), and artificially intelligent systems can do such things to greater or lesser degrees. Several benchmark frameworks specialize in evaluating the capacity of different AI models to perform some tasks related to EI, but these often lack a solid foundation regarding the nature of emotion and what it is to be emotionally intelligent. In this project, we begin by reviewing different theories about emotion and general EI, evaluating the extent to which each is applicable to artificial systems. We then critically evaluate the available benchmark frameworks, identifying where each falls short in light of the account of EI developed in the first section. Lastly, we outline some options for improving evaluation strategies to avoid these shortcomings in EI evaluation in AI systems.

Feminist Epistemologies and Trusting Chatbots

Alicia Patterson , *Libby Southgate* Epistemological Issues in AI and Computing

Generative AI, especially chatbots, are being incorporated into many epistemic domains such as schooling and research. Indeed, Ramón Alvarado (2023) argues artificial intelligence should be understood as an epistemic technology. In many of these domains, we're being asked to extend a kind of epistemic trust to generative AI chatbots.

How, then, should we think about trust from a feminist perspective when it comes to chatbots? Drawing on feminist epistemologies, we first turn to the literature for a richer account of competence which takes seriously the ways

in which who one is can affect the knowledge that one has, and thus offer to others. Knowers are fundamentally situated and one's situatedness affects one's contextually-specific epistemic trustworthiness. We need to judge if a person is well-placed to draw accurate conclusions in this context; we need to know what concepts they work with, the biases they have, their interpretation of the background information, and their projectability judgements. We argue there are systematic challenges in making these kinds of trust judgements because of the barriers to assessing the epistemic positioning of a chatbot. Our analysis has practical implications for how and where we use chatbots.

Peer Reviewed by AI

Anna Pederneschi Epistemological Issues in AI and Computing

This paper addresses the issue of generative AI and expertise, particularly in the context of the academic process for peer reviewed publications. The question at hand is: Should we use AI for peer review? I explore two key areas of analysis to address this question. An epistemic one, determining whether AI performs equally or better than an expert at the task of peer review. A socio-political one that investigates whether it is overall a good idea to substitute academic professionals with AI for peer reviews. Given the current overproduction of academic articles and the scarcity of peer reviewers, if AI is epistemically comparable to academic professionals, it would be beneficial to employ it. If this solution is not palatable, then peer reviewers should be compensated for their highly skilled labor, and the academic community should substantially reform the publishing system.

What is 'Meaningful Human Control' in Autonomous Weapons Systems?

Thomas M. Powers Human-Computational Systems Interaction

Abstract to follow.

An Overlooked Challenge for the "Emergent Multi-Scale Causality" Framework for Modelling Complex Systems. A Brief Philosophical Note

Oriol Roca-Martín Remote Presentations

Various research strands have converged around a set of closely connected ideas that make similar use of information-theoretic and computational complexity principles to account for and formally model emergence and causality across different scales; what I will call here the emergent multi-scale causality (eMSC) approaches. The framework is gaining popularity—e.g., notions such as "causal emergence" and "synergistic information" in science, and some prominent versions of ontic structural realism in philosophy. However, it has been noted by numerous authors that the methodological and metaphysical considerations that explicitly and implicitly underlie the modelling decisions of the framework remain unclear. In this context of concerns, after characterizing the core principles of eMSC, this article argues that the realist stance of the framework towards effective theories and their taxonomies is problematic. The main point is that, eMSC's strategy towards scientific realism has to be defended on two fronts, a local one and a global one, but the tools that the framework uses and develops can only focus on the local front, overlooking the global one. And this is something that allows for a multiplicity and relativity of ontologies that the realist stance of the framework wants to prima facie avoid. eMSC has two options (which can be combined). First, defend and/or develop a version of scientific realism that could be squared with a certain degree of multiplicity and relativity of ontologies. Second, explore some research avenues coming from the broader computational modelling field. By raising these concerns, some possible extensions for eMSC are noticed, while at the same time, those

computational modelling research avenues are connected with the genuine philosophical questions that eMSC is concerned about.

The Philosophy of Learned Operators: a New Kind of Scientific Model

Conor Rowan Special Topic: Epistemology of ML

Engineering practice has historically relied on models grounded in ordinary and partial differential equations (PDEs) to relate system inputs, such as forces, fluxes, or heat sources, to observable outputs, such as displacement, concentration, or temperature. These physics-based models are interpretable and generalizable, but they require deep domain knowledge and significant computational resources. Scientific machine learning (SciML), and in particular operator learning, offers an alternative strategy to predictive modeling: neural networks trained directly on data to approximate the input-output relationship traditionally furnished by PDEs. These learned operators can be highly accurate, computationally efficient, and do not require mechanistic insight into the system under study, but they lack interpretability and fail to make meaningful predictions when queried outside the training data.

In this paper, we argue that learned operators constitute a novel kind of scientific model, and are under-theorized from a philosophical perspective. Unlike traditional phenomenological laws, they attempt to entirely replace governing PDEs with high-dimensional data-driven mappings. Drawing on work in the philosophy of scientific explanation, laws, and models, we examine how surrogate models fit—or fail to fit—within existing philosophical frameworks. We argue that their limited scope, opacity, vulnerability to adversarial attack, and neglect of unobservable entities distinguish them from theory-based models in epistemically significant ways. Additionally, we suggest that new frameworks for verification and validation are required if learned operators are to be safely integrated into scientific and engineering practice.

When Nothing Needs to Be Said: Silence, Presence, and Close Personal Relationships with AI

Oluwaseun Sanwoolu Human-Computational Systems Interaction

Debates about close personal relationships with artificial intelligence often focus on responsiveness, availability, and conversational fluency as markers of intimacy and companionship. In this paper, I argue that this focus overlooks a crucial feature of close personal relationships, which is the capacity for meaningful silence. In human relationships, silence is not merely the absence of communication but a relational competence, one that can express trust, comfort, attunement, and mutual recognition. Knowing when not to speak is often as important as knowing what to say. I contend that contemporary AI systems, particularly large language models, are structurally ill-suited to participate in this form of relational silence. Designed to maximize responsiveness and output, such systems treat silence as a failure of interaction rather than as a normatively appropriate response. This limitation, I argue, reveals a deeper constraint on the kinds of relationships AI systems can sustain. While AI can simulate conversational engagement, it lacks the evaluative capacities required to recognize when silence itself is the fitting mode of presence. Thus, in this paper, I treat silence as an interactional and normative phenomenon and show that prevailing models of AI–human interaction rest on an impoverished understanding of relationality, one that equates intimacy with continuous responsiveness. Recognizing silence as a relational competence clarifies why certain forms of closeness remain out of reach for artificial systems and helps recalibrate ethical and design claims about AI companionship.

Epistemological Concerns of Missing Data

Chelsea Schwartz Ethical and Epistemological Issues in AI and Computing

This paper explores the epistemological repercussions associated with incomplete datasets in machine learning and subsequent artificial intelligence applications. Extensive literature has emerged in the last decade showing how data-driven computing technologies, such as machine learning, tend to reproduce structural inequalities due, in part, to biased datasets used in their training. This bias and its implications have often been explored from the analysis of constituent items in data sets—for example, the predominant representation of one group over another in training and benchmark datasets. This paper explores a similar yet distinct issue in dataset curation at the center of machine learning technologies: the implications of missing data. In particular, this paper focuses on how such omission may fail to encode significant cultural values into the computing architecture of AI and considers the epistemological challenges this poses, referencing what Sabina Leonelli calls data imaginaries (2021). Through this analysis, this paper argues that missing data is not simply a technical failure or oversight but an ethical and epistemological problem to be addressed. This paper helps elucidate and articulate how examining where and why missing data leads to unique forms of cultural loss perpetuated—or at least uniquely exacerbates through scale, prowess, and speed—data-driven computing technologies, such as machine learning, as they occupy more and more social spaces.

"If I Only Had a Heart": A Plug-In Emotion Variable for Consciousness-Based Models

Robert Marc Schwartz Computation, Cognitive Science, and Cognition

Consciousness research employs diverse affective scales, yet current measures remain fragmented and weakly tied to explicit consciousness models. The States of Mind (SOM) model addresses this gap by applying a reflexive self-other architecture that yields a discrete affect variable with empirically constrained balance points. By modeling positive (1) and negative (0) states of self and other across three reflection levels, SOM generates balance point states distinguishing psychological functioning from dysfunctional to optimal. Clinical and cross-cultural validation studies show observed affect balances cluster at predicted points across diverse populations and contexts. SOM's architecture aligns naturally with Active Inference accounts of consciousness, particularly Seth's formulation of self as an embodied, hierarchical system. By quantifying positive response likelihood, SOM functions as an affective readiness-to-respond variable modulating policy selection in predictive processing terms. Shifts in affective context can move SOM among balance point states, altering predicted probabilities while maintaining core consistency of the self-generative model. As a Bayesian prior, SOM establishes expectations for actions and outcomes, enabling quantitative assessment of prediction error and signaling when affective recalibration is warranted. SOM is proposed as a plug-in affect metric for consciousness-based architectures, including Active Inference, Global Workspace models and affect-modulated robotic control systems.

Separating Use from Usability in the Individuation of Computations

Mathew Smith Philosophy and History of Computing

In this paper, I develop the distinction between the problems a computation can be used for and the problem a computation is used for, if it is used at all. I argue that the identity of an algorithm is determined solely by its rule structure, independent of any intended or realized use. Whether an algorithm can be used successfully for a particular problem is an extrinsic fact relating the algorithm and the problem, not an intrinsic feature of the problem. I go on to develop what this usability relation looks like: an algorithm is usable for a problem when there exists a correct and honest coding that translates between the problem and the algorithm's computational

states. With the use/usability distinction in hand, I show how it exposes issues in arguments for extrinsic accounts of physical computation and how it clarifies computational explanations on an intrinsic account. In particular, I object to Shagrir's master argument—which exploits the phenomenon of simultaneous implementation to advance extrinsic accounts—by arguing that the extrinsic context Shagrir appeals to does not individuate a computation at all; rather, it individuates the coding used to translate between the system and its context. Finally, I provide a straightforward account of computational explanations made available to intrinsic accounts by usability: a system's performing a task is explained by the fact that it performs a computation that is usable for that task. Importantly, this account captures computationally relevant differences in how two systems perform the same task.

Hallucination as a Property of Processing Mode

Lev Sukherman Mind and Machines

Large language model (LLM) hallucinations are typically treated as architectural defects requiring technical fixes. We argue instead that hallucinations are a systemic property of the \emph{syntactic processing mode}, in which a system manipulates symbols without access to their semantic content. Our analysis focuses on base autoregressive architectures; hybrid systems incorporating explicit semantic representations or external verification may partially mitigate the effects we describe, but do not eliminate the underlying syntactic processing mode in the core generation mechanism. This perspective connects Searle's Chinese Room argument with Harnad's symbol grounding problem and empirical findings from cognitive psychology. We demonstrate that humans operating in syntactic mode produce errors that share key properties with LLM hallucinations: high confidence, surface plausibility, and reliance on pattern-matching over verification. In the Wason Selection Task, approximately 10% of subjects arrive at the logically correct answer, yet the majority report high confidence in their incorrect responses, much like LLMs producing false statements with high probability scores. Cheng and Holyoak show that accuracy increases progressively as semantic grounding moves from full abstraction to goal-directed concretization. Cosmides and Tooby extend this gradient to its furthest point: when the same logical task is recast as a social contract, human accuracy jumps to approximately 75%, revealing a capacity for mode-switching that LLMs lack. This reframes the hallucination problem: both humans and LLMs hallucinate when processing symbols without semantic grounding, and it is the processing mode, rather than the agent, that is the source of error.

Evaluating Existential Risk from AI

John Symons Epistemological Issues in AI and Computing

Abstract to follow.

From Artifacts to Agents: Socializing the Epistemology of Evidence in Law

Alejandro D. Tamez Epistemological Issues in AI and Computing

Courtrooms are paradigmatic formal epistemic environments: they are structured by evidentiary rules, institutional roles, and procedural norms designed to protect the integrity of fact-finding. Yet courts are also social institutions and thus vulnerable to the same malicious epistemic interventions that destabilize less formal social epistemic environments. As generative media systems proliferate, courts will increasingly confront not only fabricated audiovisual artifacts ("deepfakes"), but also the broader erosion of epistemic trust in digital media that has long functioned as a gold-standard evidentiary artifact. The problem is therefore not merely misinformation, but the loosening of epistemic norms as trust breaks down (Rini, 2020).

Legal scholarship has responded by emphasizing (i) provenance and forensic practices within existing doctrine, (ii) expanded judicial gatekeeping over authenticity, and (iii) stronger sanctions and professional-responsibility constraints on opportunistic "deepfake defenses" (Pfefferkorn, 2020; Delfino, 2023, 2024; Dalal et al., 2024). This Article argues that these proposals must be integrated within a social-epistemic approach to adjudication. In addition to provenance-based indicators (chain of custody, metadata integrity, capture and transfer history), judges should adopt structured, procedurally constrained assessments of agent-centered reliability: the credibility and incentives of sponsoring parties and counsel, their diligence in preserving provenance, consistency with the broader record, and litigation conduct. A two-track protocol—artifact-centered and agent-centered—aims to reduce both false admissions of fakes and strategic discounting of authentic evidence while preserving adjudicative legitimacy under deepfake uncertainty.

Algorithmic Elaboration and Agency

Caroline Wall Human-Computational Systems Interaction

Some forms of human-algorithm interaction result in the suppression of human agency. Slot machine players in Vegas, for example, describe their experiences in terms of feeling "paralyzed" or "hypnotized" by the machines' loops, to the point of not being able to get up to use the bathroom. Nevertheless, by distinguishing two structures of human-algorithm interaction, I aim to show how algorithms can either suppress or augment our agency.

I define an algorithm as any fully automated procedure for transforming an input into an output—meaning, a process involving just the abilities to read and write symbols and move between functional states. An algorithmic decision system (ADS) is a system of algorithms executable with just these abilities. When a human is incorporated into an ADS, she uses just this subset of her abilities; in doing so, she acts "automatically" or "mechanically," and her role in the ADS is substitutable.

What I term a human-algorithmic decision systems (HADS), in contrast, is a process that can be algorithmically decomposed into a set of subprocesses that includes at least one discursive practice-or-ability. As Brandom (2008) demonstrates, this necessitates the involvement of discursive creatures. I conclude by discussing aviation checklists as an example of HADS.

Can We Please Say What We Mean By Program?

Nick Wiggershaus Remote Presentations

Computer Programs are one the most central entities of computing. Yet, it remains unclear what kind of thing we are talking about. Are programs best understood as mathematical entities, physical processes, linguistic constructions, or social practices? In this talk, I offer a conceptual diagnosis for the ongoing metaphysical ambiguity surrounding such entities. Specifically, I argue that a central reason why we struggle to pin down their ontological nature lies in their unclear definition: a closer look at the etymology of programs reveals that their polysemic character is grounded in the pluralistic nature of computer science qua discipline, including traits from mathematics, the empirical sciences, and engineering. While the resulting linguistic ambiguity is largely innocuous, I submit that it is one of the root causes for why we historically and presently struggle in our metaphysical inquiries of programs. Ontological debates about whether they are abstract or concrete tend to stumble upon unnoticed shifts in reference. What looks like a disagreement about metaphysics can often be deflated to a conflation of referents under a shared label.

Convergence Without Answerability: Why Representational Alignment Does Not Constitute Epistemic Agency

Mohammad Mohsen Yasin Remote Presentations

The Platonic Representation Hypothesis (PRH) demonstrates that neural networks converge toward structurally isomorphic representations of an underlying reality, establishing what we term representational applicability — the conditions under which normative standards of accuracy and coherence apply to a system's internal states. This paper argues that applicability is not directedness. A normative space applies to a system when its representations can be evaluated against epistemic standards; but it is directed at a system only when that system stands as a source of justificatory demands — an entity answerable for its representational states, not merely assessable against them. We demonstrate this through a case of two ideal optimization systems achieving full representational convergence with a latent structure Z, while remaining constitutively incapable of normative standing: their failures are malfunctions, not delinquencies; they admit of calibration, not reproach. This asymmetry holds regardless of scale, self-monitoring capacity, or higher-order reliability. We conclude that no iteration of representational reliability yields normative standing, because standing is not a scalar property of representation. The question of which systems bear epistemic responsibilities therefore requires an independent normative theory of sourcehood — one the PRH, by design, does not provide.

Workshops

IACAP Book Club–Speakers: Ramon Alvarado, Hajo Greif, Ben Recht

Nico Formánek Workshop

The purpose of the IACAP book club is to provide you with short reviews of current books in the philosophy of computing so you don't have to read them - or make an informed decisions to do so! The book club aims to cover recent works but the occasional classic might thrown in the mix.

Reports on Courses in the Philosophy of Computing–Speakers: Ramon Alvarado, Nico Formanek

Robin K Hill Workshop

A panel discussion on teaching the philosophy of computing as a college course will both offer and solicit ideas for subject matter, materials, exercises, and descriptions, both in the ideal as suggestions and in practice as experiences. We invite IACAP attendees to share their experiences and suggestions. This workshop continues the exchange that started with a similar panel at IACAP 2023. In addition to pedagogy, we seek reflections on curricular and administrative matters, including how to situate a course in the philosophy of computing across separate academic divisions and how to reconcile the disparate background of computing and philosophy students. Our aim, especially for future computer science professionals, is to complement and expand the formal picture delivered to students. A class introducing such concepts within a solid framework of analytic philosophy will equip students to participate in the development of the field, and this workshop will help to equip interested instructors to teach those concepts.

Workshop: Artificial Wisdom – Making AI More Human and Humans Less Mechanical (Duration 60 mins)

John Sullins , *Alejandro D. Tamez* Workshop

The recent history of LLMs has shown that these systems can produce output that their users find intelligent and useful but wisdom in their outputs is encountered less often. Philosophers tend to define wisdom as a product of experience, well founded knowledge to which proper judgement is applied to produce beneficial outcomes. It can be argued to be an essential component of thoughtful ethical actions as well as a key component in working through complex situations that require discernment and a balanced approach. It is concerned with long term outcomes and is not focused entirely on short-term gain. It is the virtue we employ to foster a good and just life or philosophical eudaemonia. Cultivating this skill was a concern of ancient philosophers and has become of renewed interest to some modern philosophers. There is a profound difference though, unlike our ancient collogues, we routinely think with and through machines. These days that includes LLMs which even act as conversational partners though which we can engage in something like a Socratic dialogue. In this workshop our presenters will make a position statement that addresses some of the deep philosophical questions in this topic. Are machines capable of helping us grow through wise counsel, or do they only produce simulacral glossolalia that is attractive and convenient but which we mistakenly accept as advice that is just and good? Our panelists will explore methods such as artificial wisdom, artificial phronesis, and other methods that help ensure that the AI systems that are being designed now will fare better in their interactions with human users.

Abstracts of the Posters

IACUCs for AI: A Call for Experimentation

Bob Fischer, *Taylor Matalon* Ethics of AI, Computation, Information, and Robotics

While it seems unlikely that current AI systems have morally relevant interests, these systems are evolving rapidly enough to generate some concern (Butlin et al. 2023). Out of an abundance of caution, therefore, it's worth identifying strategies for mitigating the ethical risks of research on these systems (Birch 2024). The Institutional Animal Care and Use Committee (IACUC) model offers a potential template. IACUCs are the primary oversight mechanism for research involving animals in the United States, emerging from the US Animal Welfare Act and codified in Public Health Service (PHS) policy (OLAW 2015). These committees consist of at least five members—including a veterinarian, practicing scientist, non-scientist, and unaffiliated community representative—who review proposed research protocols based on the 3Rs. The 3Rs are the cornerstone of animal research ethics in much of the world, requiring researchers to replace live animals with non-animal models when feasible, reduce the number of live animals used (if animals must be used), and refine procedures to make them less aversive (Russell & Burch 1959). This paper considers the prospects for adapting the IACUC model for AI research oversight, flagging some of the conceptual and practical challenges and suggesting possible ways of navigating them.

Algocracy, Domination and AI

James McCollum Societal and Environmental Impact of Computing Technologies and Automated Systems

The last twenty years has seen the rise of digital platforms and their associated power structures. Digital platforms are the infrastructure of market, informational, and social interactions, so they can set both visible and invisible constraints on these relationships. These constraints have been called "algocracy". Given this relatively new form of social organization, political philosophers and theorists have sought to evaluate and conceptualize algocracy in the terms of contemporary political theory. Are digital platforms dominating systems or do they merely enable individuals? Dorothea Gädeke has argued that "systemic domination refers to systematic disempowerment in situations in which a disempowered person does not face a particular dominator with an actual capacity to interfere." While systemic domination may enable interpersonal domination, it is a locus of domination on its own and an important one for critical social analysis. I will argue that platform power should be conceived of in this way. Digital platforms are loci of systemic domination, and AI platforms may become even more intense forms of systemic domination since they will be more autonomous and less accountable. Moreover, we need a systemic conception of domination when the attending disempowerment will be pervasive and not necessarily controlled by human agents.

A Pilot Study Evaluating Emotional Intelligence in Multi-Turn Dialogue with AI Systems

Max Parks, *Kheli Atluru* Human-Computational Systems Interaction

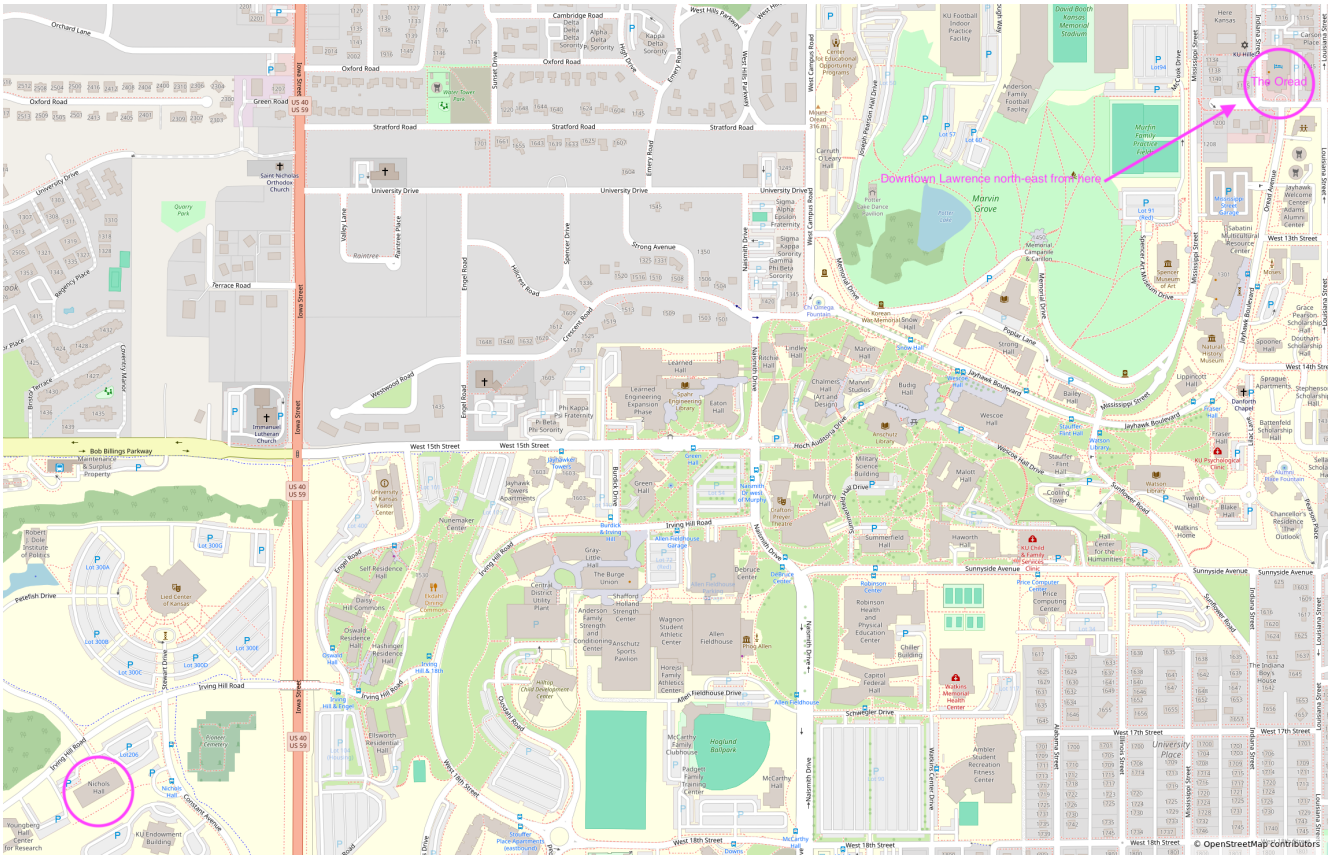
We present an IRB-pending pilot benchmark of emotional intelligence (EI) in AI chat systems using short, multi-turn dialogues rather than single prompts. Responses are rated on Sense (emotion cues without overreach), Explain (contextual appraisal with calibrated uncertainty), and Respond (empathic, autonomy-supportive, non-

manipulative guidance), and summarized with a weighted composite emphasizing Respond. The dataset includes 75 prompt/response sets ($\geq$3 turns). Two frontier models were run and logged with promptfoo; a third model not supported by promptfoo was tested manually using the same prompt sets. Psychology-trained undergraduates (BA+ or in-progress) will score each dimension on a 0/0.5/1 rubric, completing 25 items per sitting with optional additional batches. Preliminary checks suggest differences in conversational follow-up questions, emoji-based warmth, and overconfident jealousy inferences. Multi-turn testing also exposed harness artifacts: repeated relationship-term variants triggered meta-comments about "testing" and, in at least one case, refusal to continue. Next steps include testing different settings (e.g., warmth), multilingual text-only expansion, larger datasets, and multimodal evaluation.

Practical Information

Conference Venue

The IACAP 2026 Conference will be held on the campus of the University of Kansas, in Lawrence, Kansas. Conference sessions will take place inside Nichols Hall at the headquarters of the Center for Cybersocial Dynamics (see the building encircled in pink at the bottom left of the map below).



The Oread Hotel is roughly a 30+ minute walk from our conference venue (in the upper right corner of the map above, straddling Downtown) or a ~20-minute slow ride through campus on public transportation (e.g., route 45). If you are coming from downtown Lawrence, you could be at our venue via a 25+ minute bus ride (e.g., on route 4).

Technical Infrastructure

All rooms have full audiovisual equipment and are wheelchair accessible.

Wifi is available as eduroam and via the KU Guest open wifi network.

Communication Channels

Please chat with other attendees and IACAP members about this conference on the IACAP Slack Workspace:

https://iacapconf.org/slack.

Conference Social Events

Welcome drinks (self-paid) will be had on Wednesday, July the 15th at 7:00 PM at The Nest, the rooftop of the Oread Hotel on campus.

Conference Dinner

Thursday, July the 16th from 5:00-7:00 PM at Juniper Hill Farm, right outside of Lawrence. Menu will accommodate vegans and vegetarians. It includes and open bar.

https://www.juniperhillfarmtable.com/

We will take a bus that meets at Nichols Hall at 4:30 PM.