

# EPISTEMOLOGY OF AI CONFERENCE

SATURDAY FEBRUARY 7, 2026

8:30 AM – 6:30 PM

HUMANITIES GATEWAY

1010 & 1030



## Speakers:

Annalisa Coliva, UC Irvine

Peter Graham, UC Riverside

John Greco, Georgetown

Nikolaj Pedersen, Yonsei University, UIC

Yunlong Cao, UC Irvine

Tori Helen Cotton, UC Irvine

Edward Mark, Loyola Marymount University

Anna Pederneschi, UC Irvine

**UC Irvine**

Humanities Center

**UC Irvine**

Department of Philosophy

## **Schedule:**

**8:30–8:50 AM**

**Registration**

**8:50–9:00 AM**

**Opening Remarks**

**9:00–10:10 AM**

**John Greco, Georgetown University**

**“Coming to Know from Generative AI:  
Several Alternative Models”**

**Abstract:** In previous work I have proposed several models for understanding how generative AI might be understood as generating and/or transmitting knowledge. These include the External Instrument Model, the Artificial Agent Model, the Extended Agent Model, and the Group Agent Model. This paper further articulates these alternatives for understanding the relevant phenomena and considers some of their strengths and weaknesses. It emerges

**that a pluralistic approach might be the best option, with different models working best for different technologies.**

**10:10–10:30 AM**

**Coffee Break**

**10:30–11:40 AM**

**Annalisa Coliva, UC Irvine**

**“From hinge trust to e-trust”**

**Abstract: Drawing on a Wittgenstein-inspired account of trust developed in Coliva (2025), this paper examines the conditions under which it is meaningful to speak of trust in artificial intelligence (AI), or e-trust. It argues that prevailing interpersonal accounts of trust encounter serious difficulties in explaining the very possibility of e-trust, and that they often conflate trust with trustworthiness. By contrast, the paper**

**shows that the notion of hinge trust provides a more adequate framework for understanding how trust in AI is possible in the first place.**

**11:40 AM–12:20 PM**

**Anna Pederneschi, UC Irvine, School of Humanities**

**“Peer Reviewed by AI”**

**Abstract: 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.**

**12:20–1:50 PM**

**Lunch**

**1:50–2:30 PM**

**Ted Mark, Loyola Marymount University**

**“Oratio Obliqua and the Grammar of Large Language Models”**

**Abstract: This paper develops a theoretical framework for understanding the role of large language models (LLMs) in the process of knowledge acquisition. I argue that LLMs are systems that are at least theoretically capable of reliably generating indirect reports—specifically, reports about what some ‘speaker’ E would say in response to a given prompt. On this model, an agent S is justified in believing a proposition p on the basis of an LLM’s output if the following conditions are met: (i) speaker E is an epistemic authority relative to a domain in which p is a relevant proposition; (ii) there exists an utterance in which E asserts that p; (iii) the LLM reliably reports the semantic content of E’s**

assertion that  $p$ ; and (iv)  $S$  justifiably believes that the LLM is a reliable reporter of  $E$ 's assertion. I conclude by explicating a few of the practical ramifications of this account and noting some of the limitations of its application.

**2:30–3:10 PM**

**Yunlong Cao, UC Irvine, Philosophy**

**“Mechanistic Transparency of Computer Programs”**

**Abstract:** This paper argues that computers, including contemporary AI systems, are mechanistically transparent and therefore should not be attributed mental properties. I extend the concept of transferability—the property that mathematical proofs can be verified through inspection alone—to computer algorithms generally.

Using the Curry-Howard correspondence and/or the Church-Turing thesis, I show that relevant experts can in principle verify any computer algorithm's correctness by examining its steps. This transferability entails mechanistic transparency: the algorithm itself provides the best explanation for a computer's output, making mental explanations unnecessary. I apply this to two cases where mental attributions seem tempting: particular outputs and general capacities. I address objections, including concerns about AI opacity, by distinguishing mechanistic transparency from other notions of algorithmic transparency. The upshot: because of the mechanistic transparency, no inference to the best explanation supports attributing consciousness, intelligence, or intentionality to computers.

**3:10–3:50 PM**

**Tori Cotton, UC Irvine, LPS**

**“Digital Dirty Laundry: Conversational AI and the Epistemic Value of Unguarded Data”**

**Abstract: Popular conversational AI tools such as ChatGPT, Gemini, and Claude combine large language models with conversational interfaces. These systems are trained on massive datasets that include formal writing, indexed webpages, and spontaneous personal disclosures, which are later fine-tuned through user interaction. Because much of the language they process is unguarded, their outputs often reflect how people speak when they are not actively managing how they wish to be perceived. Drawing on standpoint epistemology, I argue that LLM-chatbots occupy a doxastic position analogous to that of insider-outsiders: marginalized people who are present but overlooked.**

**I describe LLM-chatbots as occupying unique doxastic positions—not as conscious agents, but systems trained to take on belief-like informational states based on patterns in human data. Although they lack the lived experience typically required for traditional standpoints, their outputs may reveal unique social insights often obscured in more curated or self-conscious forms of communication. Ignoring these insights risks overlooking the epistemic value that can emerge from such content.**

**3:50–4:10 PM**

**Coffee Break**

**4:10–5:20 PM**

**Peter Graham, UC Riverside**

**“Did Claude Tell You That? Cappelen and Dever on Chatbots and Artificial Speech Acts”**

**Abstract: Cappelen and Dever provide a prima facie case and an argument for the claim that Chatbots like Claude and ChatGPT have minds and perform full-blown speech acts. I present and examine their case.**

**5:20–6:30 PM**

**Nikolaj Pedersen, Yonsei University, UIC, Seoul**

**“Is AI Safe for Knowledge?”**

**Abstract: This talk investigates epistemological issues raised by the use of AI in enquiry: under what conditions do human AI-based beliefs qualify as knowledge (Q1), and do the seemingly crazy errors that AI systems sometimes make—including image misclassifications triggered by humanly imperceptible pixel manipulations—**

**pose a threat to AI-based beliefs qualifying as knowledge (Q2)? These questions are discussed from the point of view of modal epistemology—more specifically, through the lens of the safety condition on knowledge. It is argued that many true AI-based beliefs do not qualify as knowledge because they violate safety. Empirical and conceptual reasons are given in support of this claim, drawing on the large literature on adversarial examples and AI image classifiers.**

**6:30–8:30 PM**

**Dinner**