CHAPMAN UNIVERSITY | Schmid College of Science and Technology

AI EPISTEMOLOGY

Workshop at Chapman University

Friday, February 6, 2026
Killefer School Conference Room A

8 - 9 a.m.	***Breakfast***
9 - 10:30 a.m.	**SILVIA DE TOFFOLI** IUSS Pavia The Technological Turn in Mathematics (joint work with Fenner Tanswell)
10:30 - 10:45 a.m.	***Break***
10:45 a.m. - 12:15 p.m.	**GIORGIO GHELLI** University of Pisa Internals of Conversational Agents: Anatomy of ChatGPT
12:15 - 1:30 p.m.	***Lunch***
1:30 - 2:45 p.m.	**ALEXANDER KURZ & JONATHAN WEINBERGER** Chapman University Mathematics Engineering vs Software Engineering
2:45 - 3 p.m.	***Break***
3 - 4:15 p.m.	**ALESSIO TACCA** IUSS Pavia Appropriate Reliance on AI Systems
4:15 - 4:30 p.m.	***Break***
4:30 - 5:45 p.m.	**URI MAOZ** Chapman University Understanding the Intentions of AI Agents

SILVIA DE TOFFOLI

IUSS Pavia

The Technological Turn in Mathematics (joint work with Fenner Tanswell)

Abstract: Quickly evolving technologies, such as Interactive Theorem Provers (ITPs), Automated Theorem Provers (ATPs), and Large Language Models (LLMs), all falling under the general heading 'AI for mathematics,' are transforming mathematical practice in profound ways. This talk explores the implications of these innovations, focusing on their impact on how mathematical knowledge is created and shared. It also discusses how they are reshaping the social dimension of mathematics, altering collaboration dynamics, trust relationships, and the collective production of knowledge. For instance, tools like ITPs facilitate large-scale collaborations and make new types of teamwork possible, where trust is not a necessary ingredient. ITPs also help us mitigate our human fallibility, yet they raise questions about the nature of formalization and the relationship between traditional and formal mathematics. Yet, technologies such as LLMs are reshaping the division of epistemic labour between humans and machines and urge philosophers of mathematics to ask questions about the value of their work.

GIORGIO GHELLI

University of Pisa

Internals of Conversational Agents: Anatomy of ChatGPT

Abstract: Large Language Models have transformed natural language processing and conversational AI, raising new technical and philosophical questions about intelligence and understanding. This talk offers a technical introduction to the internal mechanisms of LLMs, focusing on pre-training, alignment, and the geometric structure of learned representations. We analyze how conversational behavior emerges from next-token prediction and optimization, and how a language model is refined and is aligned to value expectations. The talk concludes with a critical assessment of LLMs, highlighting both their power as engineering artifacts and their fundamental differences from human cognition.

ALEXANDER KURZ & JONATHAN WEINBERGER

Chapman University

Mathematics Engineering vs Software Engineering

Abstract: We will argue that mathematics and software engineering are converging, possibly to the extent that they will become indistinguishable disciplines in the future. Of course, mathematics and engineering have been interacting for a long time. We call these interactions superficial to distinguish them from the novel, profound interactions that have been emerging recently, in particular in the light of computer proof assistants and AI.

ALESSIO TACCA

IUSS Pavia

Appropriate Reliance on AI Systems

Abstract: While current debates in AI epistemology often focus on whether we can trust AI systems, what it means to appropriately rely on them remains unexamined. In this paper, I develop a normative, context-sensitive account of appropriate reliance (AR) in AI-supported decision-making contexts. According to the standard view, a user relies appropriately on AI when she successfully adopts correct AI advice and rejects incorrect advice. I argue that the standard view focuses too narrowly on outcome success, ignoring crucial contextual aspects of appropriateness. Instead, I propose a new framework that describes AR in terms of fittingness to normative and epistemic standards, independent of the outcome. In my account, a user (or subject) S relies appropriately on AI advice to fulfil a task T in domain D if the following three conditions are met: (I) S has sufficient understanding of the nature, desiderata, stakes, and characteristics of T (Task understanding condition). (II) S has sufficient expertise (knowledge) in domain D to assess (A) the correctness of AI outputs in D and (B) their relevance to fulfilling T (Expertise condition). (III) S has sufficient understanding of the relevant capabilities and limitations of the AI system being used for T (AI literacy condition).

URI MAOZ

Chapman University

Understanding the Intentions of AI Agents

Abstract: AI systems are becoming increasingly capable and autonomous, yet their intelligence is profoundly different from our own. Aligning the intentions of these systems with human values is therefore a critical safety imperative. To do that we must come up with a plausible list of necessary conditions for "intention" in any system—biological or artificial. This will enable us to develop the tools to measure, characterize, and intervene on intentions in AI systems. We also need to move beyond behavioral probing and look "under the hood" by analyzing the weights and activations of the neural networks that constitute these systems. This approach not only offers a path to robust AI safety, but also promises reciprocal insights into the nature of intentions in biological systems.

The same topic will also be discussed at UC Irvine (Humanities Gateway 1010 & 1030) on Saturday, February 7, 2026 with the following speakers.

Please RSVP to this event.

9 - 10:10 a.m.
John Greco (Georgetown University): “Coming to Know from Generative AI: Several Alternative Models”

10:30 - 11:40 a.m.
Annalisa Coliva (UC Irvine): “From hinge to e-trust”

11:40 a.m. - 12:20 p.m.
Anna Pederneschi (UC Irvine): “Peer Reviewed by AI”

1:50 - 2:30 p.m.
Ted Mark (Loyola Marymount University): “Oratio Obliqua and the Grammar of Large Language Models”

2:30 - 3:10 p.m.
Yunlong Cao (UC Irvine): “Mechanistic Transparency of Computer Programs”

3:10 - 3:50 p.m.
Tori Cotton (UC Irvine): “Digital Dirty Laundry: Conversational AI and the Epistemic Value of Unguarded Data”

4:10 - 5:20 p.m.
Peter Graham (UC Riverside): “Did Claude Tell You That? Cappelen and Dever on Chatbots and Artificial Speech Acts”

5:20 - 6:30 p.m.
Nikolaj Pedersen (Yonsei University, UIC, Seoul): “Is AI Safe for Knowledge?”