Workshop: Philosophical perspectives on artificial reasoning systems

Date: October 30th 2026
Location: Emil-Figge-Straße 50, room 0.442, TU Dortmund Campus

The philosophy department of TU Dortmund is organising a one-day workshop exploring the philosophical dimensions of (limitations of) reasoning in large language models. Early LLMs were notorious for frequently generating outputs that, to users, appeared to contain basic reasoning errors. However, in recent years, researchers and AI developers have invested much effort in improving the common-sense and formal reasoning capacities of transformer-based models through reinforcement learning and similar techniques. This has led to a generation of 'large reasoning models' that perform well on various benchmarks of reasoning problems. Still, questions remain about how we should interpret these capacities. Do large reasoning models really engage in reasoning, or merely in a statistical approximation of reasoning? What accounts of reasoning are available in the philosophical literature, and are these accounts applicable to LLMs? Do LLMs fare differently with respect different kinds of reasoning (e.g. common-sense reasoning, practical reasoning or counterfactual reasoning)? What are the risks of over-attributing reasoning capacities to LLMs? This workshop brings together philosophers who work on these and related questions.

Attendance is open to anyone interested, but we would like to ask attendants to register in advance by sending an email to joris.graff@tu-dortmund.de. Registration is possible until October 26th.

09:30-10:00	Opening
10:00-11:00	Gregor Betz (Karlsruhe Institute of Technology) and Joris Graff (TU Dortmund) ***What Do We Mean When We Ask if LLMs Can Reason?***
11:00-11:30	Break
11:30-12:30	Jaroslav Malik (University of Hradec Králové) ***(Dis)continuity of Artificial Reason: On the Gap between Generality and Coherence***
12:30-13:30	Lunch break
13:30-14:30	Mirco Sambrotta (University Of Vienna) ***Can Statistical Predictions Alone Give Rise to Sapience?***
	Abstract: TBA
14:30-14:45	Coffee break
14:45-15:45	Cordelia Berz (LMU Munich) ***Pearl, Not People? Counterfactual Reasoning in LLMs***
15:45-16:00	Coffee break
16:00-17:00	Leonard Dung and Thom-Felix Thormann (Ruhr-University Bochum) ***Simulation and deceptive reasoning in LLMs***
17:00-17:30	Closing

Abstracts

Gregor Betz and Joris Graff, *What Do We Mean When We Ask If LLMs Can Reason?*
This talk provides a gentle introduction to the philosophical questions that are generated by the advent of seemingly reasoning machines. After a short introduction to the technical developments behind LLMs and more specifically large reasoning models, we will discuss different ways in which the question 'Can LLMs reason?' may be approached. We look at different philosophical views of reasoning, such as interpretationist views, computational views and social-pragmatic views, and discuss some of their implications for understanding the reasoning capacities (or lack thereof) of different types of large reasoning models. We close by drawing some broader lessons about the extent to which different conceptions of (LLM) reasoning can co-exist or whether the question 'Can LLMs reason?' should have a singular answer.

Jaroslav Malik, *(Dis)continuity of Artificial Reason: On the Gap between Generality and Coherence*
Superintelligent AI agents (SI) are typically ascribed two defining qualities: *rational coherence*, meaning their preferences satisfy decision-theoretic axioms that humans routinely violate (Yudkowsky, 2019; Russell, 2019), and *general intelligence*, meaning cross-domain capabilities that exceed human cognitive performance (Bostrom, 2014). However, these qualities appear to be in tension (Müller & Cannon, 2022). *Coherence* guides us to understand possible dissimilarities from human behaviour, whereas *generality* relies on close comparisons between humans and AIs.
I argue that the *coherence-complexity trade-off* (Thorstad, 2025; Stanovich, 2013) accounts for this tension. Studies of animal behaviour often show that less cognitively complex agents often prove more rationally coherent than humans. Broadly, the principles of rational choice are easier to follow when an organism's cognitive architecture is simpler. I argue that greater *generality* requires greater *cognitive complexity,* because disparate situations demand procedures of differing complexity, so a cross-domain-capable agent faces more processing steps and more opportunities for rational errors.
If cognitive complexity overlaps with general intelligence, then the trade-off is relevant to whether we ought to model SI behaviour in terms of *coherence* (Zhi-Xuan et al., 2025; Bales, 2023). Available data presents two classes of agents: cognitively complex but incoherent (humans), or cognitively simple but coherent (certain animals). SI agents fit neither category, creating a gap between currently known agents and supposed SI. Current AI systems, like LLMs, arguably exhibit both high complexity and systematic rational incoherence. AI theorists ought to account for how SI agents might cross this gap, or consider alternative, more continuous notions of agency.

References
Bales, A. (2023). Will AI avoid exploitation? Artificial general intelligence and expected utility theory. Philosophical Studies, 182(7), 1519–1538.
Bostrom, N. (2014). Superintelligence, Paths, Dangers, Strategies. Oxford University Press.
Müller, V. C., Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? Ratio (Oxf), 35(1), 25–36.
Russell, S. J. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.
Stanovich, K. E. (2013). Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice. Thinking & Reasoning, 19(1), 1–26.
Thorstad, D. (2025). The Complexity–Coherence Trade-Off in Cognition. Mind, 134(534), 422–457.
Yudkowsky, E. (2019). Coherent decisions imply consistent utilities. *LessWrong*. <https://www.lesswrong.com/posts/RQpNHSiWaXTvDxt6R/coherent-decisions-imply-consistent-utilities>
Zhi-Xuan, T., Carroll, M., Franklin, M., Ashton, H. (2025). Beyond Preferences in AI Alignment. Philosophical Studies, 182(7), 1813–1863.

Mirco Sambrotta, *Can Statistical Predictions Alone Give Rise to Sapience?*
This talk examines whether statistical prediction can yield conceptual competence in large language models (LLMs). It takes as its starting point Simonelli's (2026) inferentialist account of LLM sapience. Working within a Sellarsian-Brandomian inferentialism, on which sapience is conceptual understanding constituted by mastery of inferential roles, Simonelli argues that sapience can come apart from sentience. He then concludes that LLMs trained on text alone may therefore already be sapient without being sentient. The talk questions whether that conclusion is warranted. LLMs are trained to optimise predictions over token sequences, whereas conceptual competence, on the inferentialist view, requires sensitivity to how claims support, conflict with, and revise one another. It remains unclear whether behavioural evidence can distinguish this sensitivity from fluent reproduction of distributional patterns. The talk then considers whether it is worth pursuing the possibility that the shortfall lies not in the statistical character of learning but in what models are trained to learn. On that basis, it explores whether statistical optimisation could be

redirected towards inferential context, making inferential relations among declarative sentences, rather than word co-occurrence, the target of learning. Drawing on Brandom's (2008) distinction between *algorithmic elaboration* and *elaboration by training*, it investigates what such a redirection would require and whether it could provide a route from statistical prediction to conceptual competence.

Cordelia Berz, *Pearl, Not People? Counterfactual Reasoning in LLMs*
Counterfactual reasoning is central to causal explanation and increasingly relevant as large language models (LLMs) are used to support decisions. I replicate a psychological experiment by Rips and Edwards (2013), in which participants answered counterfactual questions about simple mechanical devices, with GPT models ranging from text-davinci-002 to gpt-5.6-sol. I compare the models' responses, confidence ratings and free text explanations with human judgements and with two formal accounts: Pearl's (2009) interventionist semantics and Hiddleston's (2005) minimal network theory. Across model generations, explanations become markedly more sophisticated and newer models register the difference between a component that "failed" and one that "had not operated", a cue that shapes human judgements. Yet explanatory sophistication does not bring the models closer to humans. Instead, the newest models increasingly converge on Pearl's account, while their alignment with human judgements peaks and then declines. The talk discusses what these findings do and do not show: what it would mean for an LLM to reason coherently about counterfactuals, which standard it should be evaluated against and how far results from a controlled, abstract paradigm carry over to real applications.

Leonard Dung and Thom-Felix Thormann, *Simulation and Deceptive Reasoning in LLMs*
According to an increasingly influential view, the complex, human-like behavior of large language models (LLMs) should be explained as roleplay or simulation of characters. Such notions aim at enabling intentional explanations without ascribing mental states to LLMs. In this talk, I provide a taxonomy of roleplay views structured through three dimensions: Interpretation (how can LLM behavior be best explained according to the view), Mechanism (what commitments does the view have about internal mechanisms of roleplay), and Generality (does the view aim to explain all LLM behavior or specific kinds). I then analyze, with regard to the case of LLM deception, why it is practically relevant whether LLMs are mere roleplayers or engaging in deceptive reasoning. However, I discuss two challenges for empirically distinguishing these two: Similarity (Roleplay is inherently difficult to distinguish from the real thing) and Ambiguity (it is often ambiguous what roleplay views predict). Finally, I propose three strategies for overcoming these challenges, making progress towards transforming roleplay from loose metaphor to a testable theory.