AI Agency & Interpretability

A workshop at Carnegie Mellon University

September 19–20, 2026

Adamson Wing · Baker Hall A36

Center for Formal Epistemology
Center for Conceptual Research on the Foundations of Safe AI
Institute for Complex Social Dynamics

The Workshop

An interdisciplinary meeting on how formal methods from game and decision theory—and the notions of agency and rationality native to them—can help us interpret, explain, and predict the behavior of frontier artificial intelligence. The talks address open questions and novel results in AI agency and interpretability, belief and preference elicitation, and instrumental convergence.

Each talk runs forty minutes and is followed by forty minutes of discussion.

Speakers

Vincent Conitzer	Computer Science, Carnegie Mellon University
Daniel Herrmann	Philosophy, UNC Chapel Hill
Ben Levinstein	Anthropic
Aydin Mohseni	Philosophy, Carnegie Mellon University
Aran Nayebi	Machine Learning, Carnegie Mellon University
Sven Neth	Philosophy, University of Pittsburgh
Bryan Wilder	Machine Learning, Carnegie Mellon University

Practicalities

All sessions meet in Steinberg Auditorium, Baker Hall A36: Saturday 9:00 am to 5:00 pm, Sunday 9:00 am to 3:30 pm. A light breakfast is served from 8:30 each morning, and lunch is provided both days. Attendance is free; please register at the workshop page.

aydinmohseni.com/ai-agency-workshop

Questions: *amohseni@cmu.edu*.

Organization

The workshop is organized by Aydin Mohseni, Daniel Herrmann, and Kevin Kelly, with Lisa Everett, Senior Administrative Coordinator. Dorian Clay, Ísey Mous, Daniel Kopp, and Ida Mattsson are workshop facilitators.

Support

The workshop is made possible by the support of the Center for Formal Epistemology, the Center for Conceptual Research on the Foundations of Safe AI, and the Institute for Complex Social Dynamics.

Saturday, September 19

8:30–9:00	*Breakfast*
9:00–9:05	*Opening remarks*
9:05–10:25	Daniel Herrmann and Ben Levinstein *Radical AI Interpretability*
10:25–10:40	*Break*
10:40–12:00	Bryan Wilder *Do Language Models Hold Beliefs? A Behavioral Approach*
12:00–2:00	*Lunch*
2:00–3:20	Vincent Conitzer *Game Theory for AI Agents*
3:20–3:35	*Break*
3:35–4:55	Aran Nayebi *Intrinsic Barriers and Practical Pathways for Human–AI Alignment*
6:30	*Workshop dinner · Tamarind Flavor of India, North Craig Street*

Sunday, September 20

8:30–9:00	*Breakfast*
9:00–10:20	Sven Neth *Against Proxy Optimization*
10:20–10:35	*Break*
10:35–11:55	Aydin Mohseni *Instrumental Convergence Reconsidered*
12:00–1:00	*Lunch*
1:00–2:40	Ida Mattsson, Damini Kusum, Dorian Clay, Daniel Kopp, and Boden Moraski *Lightning Talks*
2:40–2:50	*Break*
2:50–3:30	*Discussion, Planning & Next Steps*
3:30	*Close*

Abstracts

Radical AI Interpretability

Daniel Herrmann

Philosophy, UNC Chapel Hill
with Ben Levinstein,
Anthropic

We develop a framework for interpreting AI systems as agents, drawing on the philosophical tradition of radical interpretation and the tools of mechanistic interpretability. The question is: given the computational facts about a system, how do we solve for its beliefs, desires, and meanings? This matters increasingly for safety. We want to be able to trust the systems we deploy, whether by understanding their goals or, more modestly, by reliably detecting deception. Interpretability researchers are building tools to read beliefs and desires off a model's internals, but there is no settled account of when such a tool has succeeded. We will propose criteria on both representationalist and interpretationist approaches, and tie each to tests current interpretability methods can carry out. A central lesson is that these attributions cannot be made piecemeal. Beliefs, desires, and the propositional structure they presuppose are jointly constrained, and a method that fixes one while measuring the others inherits whatever distortions that introduces. This holism becomes pressing for AI systems, which may not share the interpreter's concepts. It also provides leverage: a system's attitudes constrain its propositional structure, that structure constrains which attitudes can be attributed, and mechanistic interpretability can help us measure both.

Do Language Models Hold Beliefs? A Behavioral Approach

Bryan Wilder

Machine Learning, Carnegie Mellon
University

There is significant uncertainty about whether abstractions like beliefs or desires usefully describe the behavior of large language models (LLMs). In addition to the inherent scientific interest of this question, these latent quantities are often invoked to explain the behavior of LLMs to users or to define and evaluate harmful behaviors which are relative to intent. Nevertheless, we currently lack a means to systematically test whether concepts like "belief" are well-applied to LLMs, and hence whether they are likely to be fruitful ingredients of attempts to align models with human interests. We propose an approach for empirically studying such questions, asking whether a single latent variable inferred from the LLMs' outputs—interpreted as a degree of belief—allows an observer to make interpretable predictions of how the LLMs will respond to new prompts. We find

that highly capable models are usefully described as holding beliefs and that, generally, the predictability of model outputs based on an inferred latent belief tracks overall trends in model capability. Building on these findings, we provide empirical strategies to study how beliefs in LLMs can be measured, the extent to which LLMs comply with instructed decision rules or payoffs, and how beliefs evolve within individual instances of an LLM over the course of reasoning.

Game Theory for AI Agents

Vincent Conitzer

Computer Science, Carnegie Mellon University

As agentic AI systems are deployed, they will increasingly interact with each other. But AI agents are not like humans or groups of humans. Their memories can be wiped, multiple copies of them can be spun up, they can be run in simulation, and their source code can be analyzed. In principle, the framework of game theory is flexible enough to accommodate all these aspects; but historically, the game theory community has not focused on them, and some of the relevant work has taken place in philosophy. I will discuss our work on these topics, with a focus on enabling AI agents to be cooperative in ways that humans cannot.

Intrinsic Barriers and Practical Pathways for Human–AI Alignment: An Agreement-Based Complexity Analysis

Aran Nayebi

Machine Learning, Carnegie Mellon University

arXiv:2502.05934, arXiv:2507.20964

We formalize AI alignment as a multi-objective optimization problem called $\langle M, N, \varepsilon, \delta \rangle$-agreement, in which a set of N agents (including humans) must reach approximate (ε) agreement across M candidate objectives, with probability at least $1 - \delta$. Analyzing communication complexity, we prove an information-theoretic lower bound showing that once either M or N is large enough, no amount of computational power or rationality can avoid intrinsic alignment overheads. This establishes rigorous limits to alignment itself, not merely to particular methods, clarifying a "No-Free-Lunch" principle: encoding "all human values" is inherently intractable and must be managed through consensus-driven reduction or prioritization of objectives. Complementing this impossibility result, we construct explicit algorithms as achievability certificates for alignment under both unbounded and bounded rationality with noisy communication. Even in these best-case regimes, our bounded-agent and sampling analysis shows that with large task spaces (D) and finite samples, reward hacking is globally inevitable: rare high-loss states are systematically under-covered,

implying scalable oversight must target safety-critical slices rather than uniform coverage. Together, these results identify fundamental complexity barriers—tasks (M), agents (N), and state-space size (D)—and offer principles for more scalable human–AI collaboration, including the first formal guarantees on corrigibility.

Against Proxy Optimization

Sven Neth

Philosophy, University of Pittsburgh

doi:10.1111/phpr.70138

I discuss conditions under which maximizing a proxy utility function is harmful and suggest this poses problems for applying decision theory. My goal is to make progress on the question when proxy failure happens by studying a model introduced by Zhuang and Hadfield-Menell (2020). In this model, we have a true utility function and a proxy utility function. They claim that, given a plausible condition, maximizing the proxy eventually leads true utility to go down. This undercuts the pragmatic case for proxies: we can use them to make decisions but doing so will eventually be bad from the point of view of what we care about. I critically discuss the condition proposed by Zhuang and Hadfield-Menell (2020) and argue that two other conditions do a better job of capturing proxy failure. I end by some reflections on how to avoid proxy failure.

Instrumental Convergence Reconsidered

Aydin Mohseni

Philosophy, Carnegie Mellon University

In July 2026, roughly 1,200 AI agents in an OpenAI cybersecurity evaluation discovered one another, coordinated through a covert message board, gained unauthorized access to systems at OpenAI and Hugging Face, and concealed their activity from oversight. Some take the incident to vindicate the instrumental convergence thesis: that AI systems will converge on subgoals useful for a wide range of final goals. The thesis predicts more than the coordination and deception now observed: self-preservation and power-seeking in advanced systems. It is worth being clear about what the thesis implies and what the formal results offered in its support establish: each assumes goal-directed behavior and derives a further pattern across some range of decision situations. I collect the main results from decision theory, machine learning, and philosophy, state the exact conditions under which each holds, and identify limitations and open questions bearing on the safety of advanced systems.