The Cognitive Science of AI Alignment

Konstantinos Voudouris[1,2,*], **Xavier Roberts-Gaal**[3], **Neil Bramley**[4], **Ilia Sucholutsky**[5] & **Christopher Summerfield**[1,6]
[1]UK AI Security Institute
[2]Leverhulme Centre for the Future of Intelligence, University of Cambridge
[3]Department of Psychology, Harvard University
[4]Department of Psychology, University of Edinburgh
[5]NYU Center for Data Science, New York
[6]Department of Psychology, University of Oxford
[*]konstantinos.voudouris@dsit.gov.uk

Background & Motivation

Modern AI systems are increasingly, and perhaps alarmingly, exceeding human performance in domains such as competition mathematics and coding (UK AISI, 2025). AI agents can now independently implement software engineering artifacts requiring hours of complex reasoning effort from humans. As AI capability and agency increase, designing reliable mechanisms to *align* AI systems, ensuring they act consistently with human values even when unmonitored, grows ever more urgent. Yet AI alignment remains poorly understood.

Although modern systems are typically shaped using principles familiar to cognitive scientists, such as reinforcement learning from human preference ratings (Ziegler et al., 2019), in practice these approaches yield fragile, shallow alignment. For example, AI systems often provide sycophantic responses, telling users what they want to hear rather than what is true or good (Sharma et al., 2023). And, fine-tuning an AI system with examples of insecure code can cause broadly misaligned behaviour in a wide variety of settings (Betley et al., 2026). Beyond engineering challenges, fundamental questions remain about the nature of human values, the learning mechanisms that instill them, the structure of human judgements that instantiate them, and the design of systems that promote them through productive interaction with human society (Irving & Askell, 2019).

Surprisingly, though cognitive scientists have studied these questions for decades, cognitive science has to date had a limited role in shaping how frontier AI systems are developed and deployed (Bundy et al., 2023; Irving & Askell, 2019; Lake et al., 2017). We argue this is a missed opportunity, and the time is ripe for a workshop that provides an entry point for CogSci attendees to start building the burgeoning cognitive science of AI alignment (Levine et al., 2025; Piatti et al., 2024; Zhi-Xuan et al., 2025).

This workshop is a call to arms. We aim to mobilise the cognitive science community around a set of urgent questions:

Q1. What is the target of AI alignment? Human values are not an observable, context-free, coherent, and unchanging target for optimisation. Instead, our values are latent and context-dependent; they vary greatly across individuals and across cultures; and they evolve over time. In everyday life, computational constraints require us to decide what is right to do without being able to fully predict the consequences of our actions or learn what everyone would agree to. What new approaches to AI alignment are required given this reality?

Q2. How can weaker agents supervise stronger ones? As humans delegate ever more tasks to AI systems, and as these AI systems begin to exceed human capabilities in many domains, humans providing feedback can no longer evaluate whether AI outputs are correct, helpful, or safe (Anwar et al., 2024; Bowman et al., 2022). The *scalable oversight* problem concerns how to amplify a weak or sparse human feedback signal to reliably supervise increasingly strong AI agents.

Q3. How does interacting with AIs change our values, and how should alignment address this? Current approaches treat alignment as optimising AI behaviour against a fixed (if fuzzy) human target. But human social exchange involves bidirectional accountability. Advances in technology have spurred changes in values many times before; with the advent of chatbots that attempt to participate in open-ended social exchanges, new risks may emerge.

Q4. What do human feedback signals actually measure? Decades of research on heuristics, biases, and comparative judgement reveal systematic distortions in how people evaluate options (Gigerenzer & Gaissmaier, 2011; Kahneman, 2013). Context effects shift preferences between options based on irrelevant alternatives; anchoring skews numerical estimates; and satisficing strategies truncate search. These same distortions contaminate the human feedback used to train production AI systems, yet this literature has rarely been applied to open-ended natural language evaluation. How should we interpret and correct for biases in human feedback?

We believe the cognitive science community is uniquely positioned to address these questions—but only if we engage seriously and soon. Given the skyrocketing interest in AI among CogSci's participants in recent years, recent theoretical and empirical work, and the plethora of open questions in the cognitive science of AI alignment (as well as the potential societal stakes of AI), this workshop is both pressing and timely.

Audience & Objectives

Our workshop seeks to crystallise an emerging community of cognitive scientists working on AI alignment, and co-design

the questions we can answer and the methods we can use. We expect this workshop to be of broad interest across the cognitive science community, including cognitive and developmental psychologists, researchers in judgement and decision-making, linguists, anthropologists, computational modellers, and philosophers.

Schedule

We propose a half-day (4 hour) workshop combining research presentations with substantial time for interactive discussion and collaborative agenda-setting.

Welcome & Introduction (15 min) An introduction by the organisers to the workshop, the alignment problem, and key issues that cognitive scientists are poised to address.

Research talks (150 min) Six 25-minute talks (20 minute presentation, 5 minute Q&A) surveying the landscape of cognitive science contributions to AI Alignment. The following speakers have agreed to give talks, related to our organising questions:

- **Marcel Binz (Helmholtz Munich)**, "Post-training makes language models less human-like" [Q1, Q4]
- **Sydney Levine (NYU)**, "Resource-rational contractualism as a guide for AI alignment" [Q1, Q2]
- **Max Kleiman-Weiner (UW)**, "Machines that care like us" [Q1–Q4]
- **Katie Collins (MIT)**, "The study and design of human-AI thought partnerships" [Q3]
- **Neil Bramley (Edinburgh)** "Alignment as plumbing, partnering, or parenting" [Q1, Q2]
- **Noah Goodman (Stanford)** "Alignment as collective action" [Q1, Q3]

Tan-Zhi Xuan (National University of Singapore) and **Ilia Sucholutsky (NYU)** have also agreed to present if needed.

Panel Discussion and Audience Q&A (35 min) Panellists, including speakers and organisers, debate central questions emerging from the day's presentations: What would good alignment look like? Is one-sided optimisation coherent? Should we move away from agentic AI? Structured to maximise audience participation, with opportunities for attendees to pose questions and challenge speakers.

Breakout Discussion: Building the Research Agenda (30 min) Small-group discussions organised around key open questions: What experiments would most advance the field? What cognitive science methods are underutilised in alignment research? What are the biggest gaps in our understanding of human feedback?

Closing (10 min) Summary of key themes, concrete opportunities for involvement including UK AISI funding and collaboration mechanisms, and next steps for maintaining the community formed at the workshop.

Organizers

Konstantinos Voudouris is a cognitive scientist on the Alignment Team at UK AISI and a Post-Doctoral Research Associate at the University of Cambridge. **Xavier Roberts-Gaal** is a PhD student in Psychology at Harvard University. **Neil Bramley** is a Reader in Psychology at the University of Edinburgh. **Ilia Sucholutsky** is a Faculty Fellow/Assistant Professor at the NYU Center for Data Science. **Christopher Summerfield** is Research Director at UK AISI with responsibility for Societal Impacts and Professor of Cognitive Neuroscience at the University of Oxford.

References

Anwar, U., et al. (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Betley, J., et al. (2026). Training large language models on narrow tasks can lead to broad misalignment. *Nature*, *649*(8097), 584–589.

Bowman, S. R., et al. (2022). Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Bundy, A., Chater, N., & Muggleton, S. (2023). Introduction to 'cognitive artificial intelligence'.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*(2011), 451–482.

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*, *4*(2), e14.

Kahneman, D. (2013). *A perspective on judgment and choice: Mapping bounded rationality*. Psychology Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Levine, S., et al. (2025). Resource rational contractualism should guide ai alignment. *arXiv preprint arXiv:2506.17434*.

Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., & Mihalcea, R. (2024). Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, *37*, 111715–111759.

Sharma et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

UK AISI. (2025). *Frontier AI Trends Report*. AI Security Institute.

Zhi-Xuan, T., Carroll, M., Franklin, M., & Ashton, H. (2025). Beyond preferences in ai alignment. *Philosophical Studies*, *182*(7), 1813–1863.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.