Tristan Day

Research Fellow, Cambridge Boston Alignment Initiative

Cambridge, MA. Open to relocation. | tristan.julian.day@gmail.com | tristan-day-research.github.io | GitHub | LinkedIn

Research Statement

AI alignment researcher studying model introspection, metacognition, and internal representations using mechanistic interpretability. My work focuses on when language models have latent access to information about their own uncertainty, behavior, or role-conditioned states, and how that access can be measured or elicited reliably. Current projects span confidence self-report, persona representations, privileged access between personas, activation-level analysis, and evaluation frameworks for AI-relevant cognitive properties. I bring a background across machine learning, computational neuroscience, cognitive science, and philosophy of mind, with a focus on empirical methods that connect model internals to externally observable behavior.

Research Experience

Research Fellow, Cambridge Boston Alignment Initiative (CBAI) 2026 – Present

Persona representations and model self-knowledge. Mentor: Laura Ruis (UCL, MIT); collaborator: Belinda Li (MIT, Anthropic).

- Investigating whether LLM personas reflect deeper computational differences or mainly surface-level stylistic variation, and whether personas have privileged access to persona-specific latent information.
- Designed and implemented experiments using linear probes, CKA/SVCCA/Procrustes analysis, activation patching, SelfIE/Patchscopes, and logit lens on Llama 3.1 8B and Dolphin 2.8 24B Mistral (custom PyTorch implementation on RunPod) across various experimental setups.
- Preliminary results suggest persona identity is linearly decodable from early-layer activations, with mid-layer activation divergence consistent with persona-dependent representational structure beyond surface-level style.
- Motivated by alignment-relevant questions about model self-report, role-conditioned behavior, and the reliability of claims about internal states.

Researcher / Fellow, SPAR (Supervised Program for Alignment Research) 2025 – 2026

LLM metacognition and confidence self-report. Mentor: Christopher Ackerman.

- Investigated metacognition in LLMs via novel experiment designs: fine-tuned Llama 3.1 8B to play a game where it defers to a teammate based on its internal confidence signal, which generalized to a different task requiring metacognitive access.
- Found substantial improvement in confidence–accuracy correlation with strong transfer retention to held-out domains, indicating generalizable metacognitive structure rather than surface-level heuristics.
- Layerwise probing revealed pre-existing metacognitive representations in middle layers; compared contrast directions for output entropy and stated confidence across base, instruction-tuned, and fine-tuned stages.
- Methods: fine-tuning, evaluations, activation analysis, representation comparison, model-stage comparisons, contrastive direction acquisition, steering, and ablation.
- **Output:** ICML 2026 Mechanistic Interpretability Workshop paper/poster.

AI Evaluation Frameworks and Consciousness Indicators 2026 – Present

Research collaboration, Sentient Futures Mentorship Program. Mentor: Chris Percy.

- Developing evaluation frameworks for AI-relevant cognitive properties, building on consciousness-indicator frameworks and alignment evaluation needs.
- Translating candidate properties such as agency, world modeling, self-modeling, attention, metacognition, and valenced-state representations into concrete behavioral and architectural tests.
- Structuring philosophical and cognitive-scientific criteria into empirically tractable evaluation proposals. **Status:** early-stage / planning-stage.

Lead Researcher, Connectome Harmonics 2022 – Present

International, multi-institution collaboration; originated at Neurohackademy (Univ. of Washington).

- Coordinate a distributed research team developing methods to characterize changing neural dynamics and network structure over the lifespan using the Connectome Harmonics framework.

- Built a spectral-decomposition pipeline (graph Laplacians, eigenmodes, harmonic structure) for diffusion-MRI connectivity data collected at the University of Oxford.
- Work aims to identify biomarkers of age-related cognitive decline; manuscript planned.

Selected Research Outputs

Day, T., & Ackerman, C. *Fine-Tuning Enhances Latent Metacognitive Capability in Language Models.* ICML 2026 Mechanistic Interpretability Workshop. Workshop paper and poster presentation.

Paper | Poster | Code

Day, T. *How Deep Do Personas Go? Computational Differences and Privileged Access in LLM Personas.* Work-in-progress poster, Cambridge Boston Alignment Initiative, April 2026.

Preliminary Poster | Code

Education

Portland State University — Summa Cum Laude

M.S. Computer Science, AI / Machine Learning Track.	GPA: 3.90	2023
Graduate Certificate in Computational Intelligence (Complex Systems department).	GPA: 4.00	2022
B.S. Psychology; minors in Neuroscience and Philosophy.	GPA: 4.00	2021

Honors: President's List every term.

Relevant coursework: Philosophy of Mind, Philosophy of Cognitive Science, Cognitive Neuroscience, NLP & Advanced NLP, Deep Reinforcement Learning, Artificial Neural Networks, Artificial Life.

Additional Teaching & Professional Development

Teaching Assistant, Advanced Neurophysiological Psychology (400-level) · Portland State University · 2019
Led study sessions and lectures, graded exams, and critiqued student presentations.

Conference participation: Society for Neuroscience Annual Meeting, poster presenter, 2018 and 2019; SfN Virtual Conference on Machine Learning in Neuroscience, 2019.

Technical Skills

ML & Interpretability	PyTorch, Hugging Face Transformers, linear probing, activation patching, logit lens, Patchscopes/SelfIE, CKA, SVCCA, Procrustes, PCA, representational similarity analysis, steering, ablation, fine-tuning / LoRA, model evaluation.
Scientific Computing & Infrastructure	Python, NumPy, SciPy, pandas, scikit-learn, matplotlib, Jupyter, RunPod and remote GPU workflows, GCP/AWS, Docker, Weights & Biases, PyTorch Lightning, Linux.
Mathematics & Methods	Linear algebra, spectral analysis, Bayesian and frequentist statistics, network theory, complexity science, statistical analysis, experimental design.
Computational Neuroscience & Graph Theory	Graph Laplacians, spectral graph theory, connectome harmonics, eigenmodes, random walks, NetworkX, PyGSP.
Research Communication	Workshop papers, posters, technical writing, literature synthesis, research presentations, teaching and study-session facilitation.