WILLIAM G. P. MAYNER

Mechanistic Interpretability · Computational Neuroscience

wmayner@gmail.com
github.com/wmayner
willmayner.com
/in/will-mayner
Full CV

SUMMARY

I'm a computational neuroscientist with a mathematics & computer science background and a decade of experience developing the mathematical formalism of integrated information theory (IIT), building scientific software (PyPhi; 150+ citations), and analyzing large-scale neural data. I'm now focusing on **mechanistic interpretability and AI safety**, which I believe is the most impactful use of my expertise and skills. 13 peer-reviewed publications: first-author in *PLOS Computational Biology*, *Entropy*, and *eNeuro*; co-author in *Nature Neuroscience*. Anthropic Fellows finalist (top 130 of ~5,000 applicants).

EDUCATION

Ph.D. in Neuroscience, University of Wisconsin–Madison 2016–23

Advisor: Giulio Tononi

Thesis: Integrated Information Theory: Theoretical Developments & Empirical Applications

Sc.B. in Mathematics–Computer Science, Brown University 2009–13

AI safety training: BlueDot Technical AI Safety Course & Project (2026).

INTERPRETABILITY RESEARCH

Belief manifolds, and how to steer along them April–May 2026
Independent · BlueDot Technical AI Safety Project

- Reproduced Sarfati et al. (2026, Goodfire) "The Shape of Beliefs". LLMs represent in-context learned posteriors as curved manifolds; **geometry-aware steering** along either the primal manifold (activations) or dual manifold (linear field probes) changes the model's posterior with fewer side effects than naive linear steering. I connect this work to an earlier 'geometric turn' in computational neuroscience. The writeup drew a response from the paper's first author. [blog post]
- Currently extending the method to **evaluation awareness** and natural-language tasks (translation / code-switching).

Decomposing introspection in LLMs: representation and report March–April 2026
Independent

- Decomposed concept-injection introspection (Gemma-3 12B, Qwen-2.5 32B) into separable components: **representation** (what the model encodes about an injection) and **report** (the prompt-dependent late-layer circuitry that surfaces it), explaining apparent conflicts across prior protocols. [blog post]
- Answered two open questions from Pearson-Vogel et al. (2026): the circuitry behind prompt-framing effects, and post-training's role in building it. Concurrent with Macar et al. (2026, Anthropic) and Lederman & Mahowald (2026).

EXPERIENCE

Researcher 2023–present
Center for Sleep and Consciousness, University of Wisconsin–Madison

- Lead developer and maintainer of PyPhi, the standard open-source library for IIT research.

- Advanced IIT's mathematical formalism by developing a refinement of its intrinsic information metric (Entropy paper).
- Extended the formalism to account for perception and how systems internalize the causal structure of their environment (preprint).

Graduate Research Assistant 2016–23
Center for Sleep and Consciousness, University of Wisconsin–Madison

- Designed and conducted large-scale two-photon calcium imaging experiments in mouse visual cortex in collaboration with the Allen Institute's OpenScope program, quantifying stimulus-evoked neurophysiological differentiation (eNeuro paper).
- Core contributor to the development of IIT's mathematical formalism; co-lead author of its most recent formulation, IIT 4.0 (PLoS Computational Biology paper).
- Developed and released PyPhi (PLoS Computational Biology paper).

Associate Systems Programmer 2014–16
Center for Sleep and Consciousness, University of Wisconsin–Madison

- Conceived and built PyPhi.
- Designed and implemented evolutionary algorithms evolving neural-network-controlled agents, analyzing their emergent dynamics with information-theoretic measures.

SKILLS

ML & Interpretability: PyTorch, Hugging Face, TransformerLens, nnsight, repeng, SAE Lens, einops, nngeometry, bitsandbytes, scikit-learn, Weights & Biases
Languages & Computing: Python, C++, R, JavaScript, LaTeX; NumPy, SciPy, Pandas, Dask, HTCondor, Mathematica
Mathematical foundations: Information theory, probability, causal inference, linear algebra, optimization, discrete mathematics

SELECTED PUBLICATIONS

Full list at willmayner.com/publications. ** co-first author.*

1. **Mayner, W. G. P.**, Marshall, W., & Tononi, G. (2026). Intrinsic cause–effect power: the tradeoff between differentiation and specification. *Entropy*, 28(4), 410. [link]
2. **Mayner, W. G. P.**, Juel, B. E., & Tononi, G. (2024). Intrinsic meaning, perception, and matching. *arXiv:2412.21111*. [link]
3. Albantakis, L.,* … **Mayner, W. G. P.**,* … & Tononi, G. (2023). Integrated information theory (IIT) 4.0. *PLOS Computational Biology*, 19(10), e1011465. [link]
4. **Mayner, W. G. P.**, Marshall, W., Billeh, Y. N., … Arkhipov, A. (2022). Measuring stimulus-evoked neurophysiological differentiation in mouse visual cortex. *eNeuro*, 9(1). [link]
5. **Mayner, W. G. P.**, Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: a toolbox for integrated information theory. *PLOS Computational Biology*, 14(7), e1006343. [link]
6. Findlay, G., Marshall, W., Albantakis, L., David, I., **Mayner, W. G. P.**, Koch, C., & Tononi, G. (2025). Dissociating artificial intelligence from artificial consciousness. *arXiv:2412.04571*. [link]