Anna Tsvetkov

Brown University
Department of Philosophy
Department of Computer Science

anna_tsvetkov@brown.edu
www.annatsv.github.io

Areas of Interests

AOS: Philosophy of AI (including AI Ethics) · Philosophy of Mind · Philosophy of Cognitive Science
AOC: Philosophy of Language · Logic and Formal Methods

Education

PhD in Philosophy, Brown University 2018 - 2025
Dissertation: *Human-Centered Artificial Intelligence*
Committee: Adam Pautz (Chair), Chris Hill, Ellie Pavlick, David Chalmers (NYU)

ScM in Computer Science, Brown University 2023 - 2025
Thesis: *Can Large Language Models Represent Perceptual Reality?*
Advisor: Ellie Pavlick

BA in Philosophy, Binghamton University, *summa cum laude* 2014 - 2018

Publications

Articles

Murphy, R. A., Witnauer, J. E., Castiello, S., Tsvetkov, A., Li, A., Alcaide, D. M., & Miller, R. R. (2022). "More frequent, shorter trials enhance acquisition in a training session: There is a free lunch!" *Journal of Experimental Psychology: General*

Under Review

Tsvetkov, A. (2024). "Troubles with Spatial Functionalism"

In Progress

Tsvetkov, A. & Pavlick, E. (2024). "Can Large Language Models Represent Perceptual Reality?"
Tsvetkov, A. "Molyneux's Question and Multimodal Models"
Tsvetkov, A., Trisovic, A., Thompson, N. "A Large-Scale Analysis of Ethical Considerations in AI Foundation Models" (In progress with *MIT FutureTech.)*
Trisovic, A., Tsvetkov, A., Fogelson, A., Siva, J., Lin, S., Siminiuc, D., Vellon, R.C., Zhou, D., Yuan, G., Thompson, N. "Scientific Use of Foundation Models and Democratization of AI" (In progress with *MIT FutureTech.)*

Experience

MIT FutureTech
AI Researcher, Cambridge, MA Current

NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFA)
PhD Summer School Participant, Boston, MA Summer 2023, 2024

AI Safety Student Team at Harvard (AISST)
Technical AI Safety Fellow, Cambridge, MA Spring 2023

Presentations and Comments

(∗ = invited, otherwise peer reviewed.)

'Scientific Use of Foundation Models and Democratization of AI (with Ana Trisovic)"
MIT Initiative on the Digital Economy, MIT (May 2024) *

"Comments on Umrao Sethi (Brandeis University)"
New England Workshop on Mental Things, Rhode Island College (Nov. 2022) *

"Comments on Julia Minarik (UToronto)"
Athena in Action, Rutgers University (June 2022)

"Counting vs. Timing Intertrial Intervals During Conditioning"
International Conference on Comparative Cognition, Melbourne, FL (Apr. 2018)

Teaching

(* = an online course)

Brown University
Knowledge and Reality (Spring 2022), *Full Instructor*
Consciousness (Fall 2019, Fall 2021) *Teaching Assistant*
Philosophy of Mathematics (Fall 2020*) *Teaching Assistant*
Early Modern Philosophy (Spring 2020*) *Teaching Assistant*
Ancient Greek Philosophy (Fall 2023) *Teaching Assistant*
Ethical Themes in the Contemporary American Short Story (Spring 2024) *Teaching Assistant*
Place of Persons (Spring 2021*, Fall 2024) *Teaching Assistant*

Summer@Brown Pre-College Program
Philosophy and Psychology of Happiness (Summer 2019) *Teaching Assistant*

Graduate Coursework

(* = an audited course)

In Philosophy
Perception*
Mental Representation*
Philosophy of Science
Reductionism
Introspection
Inquiry
Kant: Critique of Pure Reason
Aristotle's Psychology
Disagreement
Moral Psychology
Independent Study (with Adam Pautz)
Independent Study (with Elizabeth Miller)

In Computer Science
Artificial Intelligence
Deep Learning
Computational Linguistics
Reading and Research I (with Ellie Pavlick)
Reading and Research II (with Ellie Pavlick)
Cybersecurity Law and Policy

In Other Institutions
Hume's Treatise and Contemporary Themes in Humean Metaphysics (MIT)
The First-Person (Harvard)

Programming Competence

Languages Python
Libraries NumPy · Pytorch · TensorFlow · Transformers · matplotlib

Service

Member of Minorities and Philosophy (MAP) Chapter (2018 - 2024)
Member of the Climate Committee Working Group (2020 - 2022)
Co-organizer, Brown's Mark L. Shapiro Graduate Philosophy Conference (2022)
Co-organizer, Philosophy Department, Inclusive Teaching Workshop (2019)
Graduate Student Mentor, Brown AI Safety Team (BAIST) and MIT AI Alignment (MAIA) (2023-2024)

References

Adam Pautz
Professor of Philosophy
Brown University
Email: adam_pautz@brown.edu

Ellie Pavlick
Briger Family Distinguished Associate Professor of Computer Science
Brown University
Email: ellie_pavlick@brown.edu

Ana Trisovic
AI Research Scientist
MIT FutureTech
Email: ana_tris@mit.edu

Dissertation Abstract

Ask a question to ChatGPT and there's a good chance you'll be impressed at what it says–unless it spits out some malarkey. Part of what is so puzzling about artificial intelligence (AI) models like ChatGPT is that no one knows precisely how they work. And without knowing how the models work, it's difficult to say what, if anything, they can teach us about ourselves. Of course, we don't *need* to know how something works to use it. I drive a car, I admit, without knowing the first thing about what goes on under the hood. But if we want to know what AI can teach us about big questions concerning the mind, understanding, and ethics my dissertation makes the argument that we had better get our hands dirty and look under the hood of AI models. (Can't say the same for my car.)

How we should approach this task is as much a philosophical question as it is a technical one. We can start big and ask, "Do AI models *really* understand questions about the world? Can ChatGPT have anything like human-level language comprehension, values, or cognition?". Or we can start small and ask, "Does this particular neuron in the model respond in predictable ways to that feature"? My dissertation is a collection of chapters that does both the big and the small. I develop what I call a *human-centered artificial intelligence* research program where human minds inspire techniques to better interpret and align AI with human values, and AI, in turn, can help illuminate how the mind works and why it works the way it does. A key lesson emerging from the chapters that by designing experiments informed by philosophy and rolling up our sleeves to look under the hood of AI models, we can see what they represent—the good and the bad—and breathe new life into longstanding philosophical debates.

Chapter 1 examines whether AI represent the world like we do. In collaboration with Ellie Pavlick, we examine whether AI models like ChatGPT trained solely on text, and with no sensory input, can still

represent color and spatial relations. By developing a philosophically-informed notion of representation suitable for AI and by using probing techniques inspired by looking into the human brain, we examine the inner workings of large language models. We find that AI models have "color neurons" and "space neurons" that represent these features. By tweaking these neurons in targeted ways, we demonstrate that the neurons form subnetworks that actually drive perceptual processing—enough to trick the models into thinking red is green and left is right with our interventions. Our findings suggest that language models have genuine representations that reflect the the structure of perceptual reality like we do—*no senses required!* This has upshots for big questions about whether artificial minds can mirror aspects of human cognition—which would require showing that they have similar basic building blocks of cognition, or representations, like we do—as well as what it takes for artificial minds to understand the world.

Chapter 2 explores whether AI can be used to make progress on longstanding philosophical debates. This chapter revisits the famous philosophical thought experiment Molyneux's Question—which asks whether someone born blind could recognize shapes by sight if they suddenly gained vision—using AI. I connect language and vision models to simulate cross-modal learning, exploring whether a language model trained on text alone can, upon receiving visual input, recognize shapes it previously knew only through language. The experiment bridges AI and philosophy to provide insights into perception and learning transfer across modalities in both natural and artificial minds. I explain how AI provides new ways to overcome the limitations of empirical studies that use human subjects who have undergone sight restoration to address Molyneux's Question. The reader who is skeptical of what AI can teach us about ourselves should treat this chapter as an invitation to see what can be gained when the rubber meets the road—when philosophy and AI work together to illuminate how the mind works.

Chapter 3 shifts gears to the ethics of AI. Diving into the inner workings of AI models is a big challenge, but it's one worth doing. As we integrate AI into high-stakes areas like healthcare, criminal justice, and social media, ensuring that these systems align with human values is more important than ever. This chapter, in collaboration with MIT's FutureTech Lab, examines the ethical considerations behind the development of foundation models—large-scale AI models like ChatGPT that are at the foundational level of many applications. We analyze how fairness, transparency, bias, safety, and other ethical principles are addressed in the development of foundation models. Our framework identifies key gaps in current practices and offers solutions for promoting the development of ethical AI.

A unifying thread running through the chapters is that bridging big questions with technical science requires grappling with the intricacies of just how a scientific approach to the philosophy of AI should proceed. These chapters aim not only to explore the performance of AI models but also to highlight the importance of designing philosophically-grounded experiments that look "under the hood" of the models to understand what they can teach us—a principle deeply rooted in empirical work in computer science. The truth is that while philosophical work on AI aims to balance philosophical inquiry with empirical findings, in practice, things often end up lopsided—and you can guess which way. A healthy dose of hands-on exploration under the hood of AI models might just do us, philosophers, good.