Anna Tsvetkov

Brown University
Philosophy and Computer Science (AI/ML)

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

Employment

Princeton University, *AI Postdoctoral Research Fellow* 2025 -

Education

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

ScM in Computer Science (AI/ML), Brown University 2023 - 2025
Advisor: Ellie Pavlick

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

Publications

Tsvetkov, A. (2025). "Can We Interpret Artificial Neural Networks As Having Beliefs and Desires?" (*Under review)*

In Progress

Tsvetkov, A. "Can Large Language Models Represent Perceptual Reality?"
Tsvetkov, A. "Molyneux's Question and Multimodal Models"
Trisovic, A.*, Tsvetkov, A.*, Fogelson, A., Siva, J., Thompson, N. "Scientific Use of Foundation Models and Democratization of AI" (*Equal contribution. In progress with *MIT FutureTech*.)

Experience

MIT FutureTech
AI Researcher, Cambridge, MA 2024 - 2025

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.)

Bilkent-UNAM Philosophy of Mind Conference

Bilkent University, Ankara, Turkey (May 2026) *

‘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*
The Nature of Morality (Spring 2025) *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
Interpretability of Language Models
Cybersecurity Law and Policy
Reading and Research I (with Ellie Pavlick)
Reading and Research II & III (with Ellie Pavlick)

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)

Last Updated: May 2025