

Juan P. Cadile

jcadile@ur.rochester.edu · juancadile.com

EDUCATION

University of Rochester, School of Arts and Sciences 2024 – Present
Ph.D. in Philosophy · Advisor: Jens Kipper · Philosophy of mind × mechanistic interpretability

Iona University, School of Arts and Sciences 2019 – 2023
Bachelor of Arts in Philosophy and Computer Science; Concentration in Cybersecurity (Magna Cum Laude)

Honors & Awards: BlueDot Impact Rapid Grants (LM Introspection; CNY AI Safety), Cosmos Ventures Grantee, Dean's Honors Scholarship, Stanford's d.School University Innovation Fellow, Stanford's AI Learning Differences Hackathon 1st Place

PUBLICATIONS & PREPRINTS

Abriendo la Caja Negra: Interpretabilidad Mecanicista y los Límites de la Explicabilidad de la Inteligencia Artificial
Chapter in *Tratado de Inteligencia Artificial* (Fernando L. Depalma, ed.), Editorial Hammurabi. Forthcoming

Detecting and Steering LLMs' Empathy in Action (*arXiv:2511.16699*) 11/2025
Empathy-in-action as a linear activation direction; analysis of the detection–steering gap.

RESEARCH EXPERIENCE

Supervised Program for Alignment Research (SPAR) | Project Lead 09/2026 – Present

- Lead three projects (nine mentees): virtue-based character training, persona vs. emotion vectors, AutoBench.
- Set methodology and design evals, incl. AutoBench's philosophically grounded agency rubrics and LLM-as-judge scoring.

EigenBench Anchor Constitutions | Philosophical Reviewer (with Lionel Levine, LAISR, Cornell) 06/2026 – Present

- Reviewed and revised 24 anchors (religious, philosophical, Indigenous); authored the Aristotelian and Confucian ones.
- Designed a new anchor schema from character development and “Teaching Claude Why”; co-developed generation scripts.
- Tested blind attribution and lexical-vs-semantic overlap across constitutions.

Conceptual Complexity & Machine Learning | with Jens Kipper 10/2026 – Present

- Translating word-vector analyses of conceptual complexity into mechanistic-interpretability experiments on LLM internals.

TRIAD, University of Rochester | Graduate Researcher (with Jon Herington) 05/2026 – 08/2026

- Showed LLM judges anchor on a corrupted source: error-detection F1 falls from 0.90 (note only) to 0.25 (31 experiments).
- Built a deterministic dose-band checker from 20.3M MIMIC-IV prescriptions: recall 0.97, no false alarms, beat every LLM judge.

CAMBRIA, Cambridge Boston Alignment Initiative | Fellow 07/2026

- Interpretability and RL (ARENA curriculum); capstone on introspection fine-tuning in a 1B model.

Flourish & OpenAudit | Behavioral audits of LLMs 2025 – 2026

- Flourish: 255 scenarios across 8 virtues (incl. adversarial and multi-turn); LLM-as-judge with a human-calibration pilot.
- Judge tracked raters on empathy (ρ 0.90) but not intellectual humility (ρ 0.13): motivated single-construct validation.

Artificial Virtuous Agents: Developing Virtuous Dispositions in Agentic AI 12/2024

- Cosmos Ventures grant: continual learning of virtue signals in multi-agent systems; probes to steer moral dispositions.

TEACHING & PROFESSIONAL EXPERIENCE

Philosophy Department, University of Rochester | Rochester, NY 08/2025 – Present
Graduate Teaching Assistant (Philosophy of AI; Ethics of Technology)

- Guest lectures: “Argument Analysis,” “Autonomous Weapon Systems,” “Work and Play in the Shadow of AI.”

CNY AI Safety | Founder & Organizer of a regional AI safety community, Rochester, NY 06/2026 – Present

Tucumán Provincial Legislature & National University of Tucumán | AI Policy Advisor & Developer 06/2025 – 10/2026

- Advised on AI policy and AI use in government (Argentina); solo-built a law-consultation tool for citizens and legislators.

PRESENTATIONS

What AI Welfare Can Buy Without Consciousness 11/2026 (upcoming)

- NYU AI Safety Association | with a workshop on Structuring Formal Arguments

Structuring Formal Arguments (workshop) | AISST–MAIA Weekend Workshop, Harvard & MIT 11/2026 (upcoming)

Problems for Language Model Introspection (poster, with Nick Meyer) 10/2026

- NYU Workshop on Introspection and Self-Report in Language Models

Aligned to What? Conceptual Prerequisites for Flourishing-Oriented AI 05/2026

- University of Oxford | OCTAI, Human-Centred AI Seminar

Information Without Representation: The Detection–Steering Gap in LLMs 04/2026

- AI & Human Values Workshop | Central New York Human Values Corridor

Philosophical Problems of Algorithmic Agency 07/2025

- Escuela de Posgrado Newman, Perú

On Recommender Systems, Flourishing, and Autonomy 04/2025

- AI & Human Values Workshop | CNYHC, Cornell University
- Rocky Mountain Philosophy Conference | CU Boulder
- III Workshop Filosofía e Inteligencia Artificial | SADAF, Argentina

SKILLS & LANGUAGES

Skills: Python (advanced), PyTorch, Hugging Face Transformers, TransformerLens, R (intermediate), JavaScript (intermediate)

Languages: English (fluent), Spanish (native), Italian (conversational), French (beginner), Chinese (beginner), Latin (beginner)

Additional Training: BlueDot Impact: Frontier AI Governance, AGI Strategy & Biosecurity; HarvardX: CS50AI