Juan P. Cadile

jcadile@ur.rochester.edu

EDUCATION

University of Rochester, School of Arts and Sciences 2024 – Present
Ph.D. in Philosophy

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: Cosmos Ventures Grantee, Dean's Honors Scholarship, Stanford's d.School University Innovation Fellow, Stanford's AI Learning Differences Hackathon #1 Place

RESEARCH EXPERIENCE

Abriendo la Caja Negra: Interpretabilidad Mecanística y Regulación de IA Forthcoming 2026

- Chapter in *Inteligencia Artificial* (L. Molina Soljan, L. García Balcarce, G. Vázquez, eds.), Editorial Hammurabi.

Detecting and Steering LLMs' Empathy in Action *(arXiv:2511.16699)* 11/2025

- Studied empathy-in-action as a linear activation direction and analyzed the detection–steering gap.

Virtue Probes 03/2025

- Built activation-based classifiers to measure and steer moral dispositions in LLMs.

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

- Designed a framework for continual learning of virtue-theoretic signals in multi-agent systems.

TEACHING & PROFESSIONAL EXPERIENCE

Philosophy Department, University of Rochester | Rochester, NY 08/2025 – Present
Graduate Teaching Assistant (Philosophy of AI; Philosophy of Technology)
- Assisted with lectures, grading, and office hours for upper-level Philosophy.

Golem Lab | Remote, LATAM 01/2024 – Present
AGI Strategy & Career Course Lead
- Built and delivered applied AI curriculum for early-career builders across Latin America.

Philosophy Department, Iona University | New Rochelle, NY 08/2023 – 05/2024
Ad Honorem Teaching Assistant (PHL 332: Logic; PHL 363: Philosophy of Psychology and Neuroscience)

PRESENTATIONS

On Measuring Virtue and Wellbeing in AI Systems

- University of Oxford, Dept. of Computer Science | Human-Centered Computing Series 05/2026

Information Without Representation: The Detection-Steering Gap in LLMs

- AI & Human Values Workshop | Central New York Human Values Corridor Workshop 04/2026

Philosophical Problems of Algorithmic Agency

- Escuela de Posgrado Newman, Perú 07/2025

On Recommender Systems, Flourishing, and Autonomy

- AI & Human Values Workshop | CNYHC, Cornell University 04/2025
- Rocky Mountain Philosophy Conference | CU Boulder 04/2025
- III Workshop Filosofía e Inteligencia Artificial | SADAF, Argentina 10/2024

Creativity as Interplay

- Rutgers-Columbia Undergraduate Philosophy Conference | New Brunswick, NJ 04/2023

The Why? Machine: Computing Causation

- Iona University Honors Thesis Day| New Rochelle, NY 04/2023
- Iona University Scholars Day| New Rochelle, NY 04/2023

SKILLS, LANGUAGES, & INTERESTS

Skills: Python (advanced), Java (advanced), R (intermediate), C (intermediate), JavaScript (intermediate)
Languages: English (proficient), Spanish (native), Italian (beginner), French (beginner), Chinese (beginner)
Interests: Interpretability, AI Safety, Algorithmic Agency, Human–AI Interaction, Philosophy of AI
Additional Training: BlueDot Impact: AGI Strategy & Biosecurity; HarvardX: CS50 Python for AI