

Paul de Font-Reaulx

CONTACT

✉ paul.defontreaulx@gmail.com
🌐 pauldfr.com

RESEARCH

SPECIALIZATION Moral Psychology, Philosophy of Cognitive Science, Philosophy of AI

COMPETENCE Logic, Decision Theory, PPE

EDUCATION

PhD in Philosophy, University of Michigan, Ann Arbor, 2020-2026

Dissertation: *The Pursuit of Value: Essays on the Cognitive Science of Motivation*

Committee: Chandra Sripada (Chair), Peter Railton, James Joyce, Rick Lewis (UMich Psychology), and David Chalmers (NYU)

Certificate in Cognitive Science

Department Visitor, New York University, 2024-2026

BPhil (Master) in Philosophy, Trinity College, University of Oxford, 2018-2020

Grade: Distinction

Clarendon Scholar (Full funding awarded ~1% of graduate students)

BA in Philosophy, Politics, and Economics, Magdalen College, University of Oxford, 2014-2017

Grade: First Class Degree

Anscombe Thesis Prize (Best philosophy thesis of the year)

EXPERIENCE

Co-Founder, Sophron Research, 2026

Co-founded an independent research initiative developing third-party evaluation standards of frontier models, fiscally sponsored by the Future of Life Foundation

[AI for Human Reasoning Fellow](#), Future of Life Foundation, 2025

Developing beneficent AI tools with a cohort of software engineers and researchers

Focus on designing behavioral benchmarks for frontier models

Received extension of initial fellowship

Research Intern, GovAI, University of Oxford, 2018

Researched the impact of AI on authoritarian governance

Wrote a report that was featured as expert testimony to the House of Representatives

Commission on Human Rights

FOUNDING

GRANTS

Future of Life Foundation, 2026

\$1,476,253 seed funding for Sophron Research

Macroscopic Ventures, 2026

\$200,000 seed funding for Sophron Research

PUBLICATIONS

Peer-Reviewed

“Reward is Evidence of Value”, *Philosophy of Science*, Forthcoming.

“MoReBench: Evaluating Procedural and Pluralistic Moral Reasoning in Language Models, More than Outcomes” (Yu Ying Chiu, Michael S. Lee, Rachel Calcott, Brandon Handoko, **Paul de Font-Reaulx** et al.), *Proceedings of ICLR 2026*, [Website](#).

“DeliberationBench: A Normative Benchmark for the Influence of Large Language Models on Users’ Views” (Luke Hewitt, Max Kroner Dale, & **Paul de Font-Reaulx**), *Proceedings of IASEAI 2026*, [Website](#).

“Do Expected Utility Maximizers Have Commitment Issues?”, *Philosophy and Phenomenological Research*, 112(1):1-23, 2026.

Invited

“Motivation, Pleasure, and Valence” (**Paul de Font-Reaulx** & Chandra Sripada), *Philosophia*, Forthcoming.

“What Makes Discrimination Wrong?”, *Journal of Practical Ethics*, 5(2):105-113, 2017.

Preprints

“Prioritization of Risks from Artificial Intelligence: A Delphi Study of 272 International Experts” (Alexander K. Saeri, Jess Graham, Michael Noetel, Peter Slattery, ... **Paul de Font-Reaulx** et al.), MIT AI Risk Initiative, 2026. [Website](#).

Non-Archival Papers

“Machine Theory of Mind and the Structure of Human Values”, NeurIPS MP2 Workshop, 2023.

“Generative Theory of Mind and the Value Misgeneralization Problem”, 2023.
AI Alignment Awards, Final Prize Winner (\$5,000).

“Alignment as a Dynamic Process”, NeurIPS ML Safety Workshop, 2022.
AI Risk Analysis Award Winner (~\$1,900).

Reports

“AI: The Consequences for Human Rights: Initial findings on the expected future use of AI in authoritarian regimes”, 2018
Governance of AI Project report submitted as expert testimony to the House of Representatives Commission on Human Rights

GRANTS

Cosmos X FIRE Grant for Truth-Seeking, 2025
\$7,500 to develop DeliberationBench

Career Transition Grant, Longview Philanthropy, 2025
\$23,900 to buy out teaching for 1 semester for work on AI sentence

Short Term Residency Grant, Institute for Humane Studies, 2024
\$13,000 to visit NYU for an academic year

Early Career Grant, Open Philanthropy, 2023

\$44,770 to buy out teaching for 2 semesters for work on AI safety

SCHOLARSHIPS

Junior Fellowship, Institute for Humane Studies, 2025 (\$6,000)

Junior Fellowship, Institute for Humane Studies, 2024 (\$6,000)

Adam Smith Fellowship, Mercatus Center, 2023 (\$10,000)

Global Priorities Fellowship Renewal, Forethought Foundation, 2023 (~\$6,300)

Global Priorities Fellowship, Forethought Foundation, 2022 (~\$6,300)

Clarendon Scholarship, University of Oxford, 2018 (~\$70,000)

Erik and Goran Ennerfelt International Scholar, EGE-Fonden, 2018 (~\$14,000)

TALKS

(Selected)

“Reward is Evidence of Value”

Philosophy of Science Association Meeting, San Diego, 2026 (Upcoming)

Society for Philosophy and Psychology, Baltimore, 2026 (Upcoming)

The First Philosophy and Reinforcement Learning Symposium, UMass Amherst, 2026

“A Conflation of Valences”

NYU Valence and Motivation in AI Systems Workshop, 2026

LSE Foundations of Animal Sentience Working Group, London (Online), 2025

“Does Reinforcement Provide Evidence about Sentience?”

Eleos Conference, Berkeley, 2025

“Do We Want Technology to Do What We Want?”

IHS Junior Fellowship, Charlotte, 2025

“What Stands to Desire as Perception Stands to Belief?”

NYU Mind Seminar Group, New York, 2024

NYU Graduate Work in Progress Group, New York, 2024

Michigan Candidacy Seminar, Michigan (Online), 2024

“The Paradox of Self-Criticism”

IHS Junior Fellowship, Raleigh, 2024

“Machine Theory of Mind and the Structure of Human Values”

NeurIPS MP2 Workshop, New Orleans, 2023 (Declined due to illness)

“Reinforcement Learning as a Model of Human Evaluative Cognition”

Southern Society for Philosophy and Psychology, Louisville, 2023

“Penelope and the Drinks”

Central APA, Denver, 2023

Second Lake Como Summer School on Economic Behaviours, Como, 2022

Filosofidagarna, Lund, 2022

Rocky Mountains Philosophy Conference, Boulder (Online), 2021

Mark L. Shapiro Graduate Philosophy Conference, Brown (Online), 2021

“Alignment as a Dynamic Process”

NeurIPS ML Safety Workshop, Online poster, 2022

“Why Do We Spontaneously Cooperate?”

PPE Society Sixth Annual Meeting, New Orleans, 2022

TEACHING

Primary

Introduction to Formal Logic, University of Michigan, Ann Arbor, Summer 2024

Instructor

Teaching

Critical Reasoning, University of Michigan, Ann Arbor, Winter 2023

Assistant

Critical Reasoning, University of Michigan, Ann Arbor, Fall 2022

Introduction to Political Economy, University of Michigan, Ann Arbor, Spring 2022

Minds and Machines, University of Michigan, Ann Arbor, Fall 2021

SERVICE

Academic Service

Admissions Committee Graduate Representative, 2023-2024

Ethics Bowl Coach, 2023-2024

Philosophers’ Annual Co-Editor, 2023

Graduate Student Working Group Organizer, 2022-2023

Picnic Organizer, 2021-2022

Social Chair, 2021-2022

COMPASS Organizer, 2020-2021

Founder and President, Project Access, Sweden, 2016-2017

Refereeing

Philosophy Compass, Synthese, Moral Philosophy and Politics, Journal of Austrian Economics, Journal of Consciousness Studies, NeurIPS MP2 Workshop 2023, ICML Second Workshop on Technical AI Governance Research 2026, and submissions to the 2026 Apart Research AI Manipulation Hackathon.

REFERENCES

Chandra Sripada

Theophile Raphael Professor
Department of Psychiatry and Philosophy
University of Michigan, Ann Arbor
sripada@umich.edu

James Joyce

C. H. Langford Collegiate Professor
Department of Philosophy
University of Michigan, Ann Arbor
jjoyce@umich.edu

Brad Weslake

Associate Professor
Department of Philosophy
New York University, Shanghai
brad.weslake@nyu.edu

Peter Railton

Gregory S. Kavka Distinguished Professor
Department of Philosophy
University of Michigan, Ann Arbor
prailton@umich.edu

David Chalmers

Professor of Philosophy and Neural Science
Department of Philosophy
New York University
chalmers@nyu.edu

Anna Edmonds (Teaching)

Lecturer
Department of Philosophy
University of Michigan, Ann Arbor
annaedmo@umich.edu