Jacqueline Harding

hardingj@stanford.edu | jacquelineharding.github.io

AOS Philosophy of AI (including AI Ethics)
AOC Logic and Formal Methods, Philosophy of Mind

EDUCATION

Stanford University

PhD Philosophy and Symbolic Systems (2020 -- present)

Institute for Logic, Language and Computation (ILLC), University of Amsterdam

MSc Logic and Computation (2019), *Cum Laude (GPA: 9.1/10)*

Trinity College, University of Cambridge

MPhil Philosophy (2017), *Distinction*
BA Philosophy (2016), *First Class with Distinction ("Starred First")*

NON-ACADEMIC RESEARCH EXPERIENCE

Anthropic

Core Contributor, AI Moral Patienthood Project (December 2023-April 2024)

Center for AI Safety (CAIS)

Research Fellow (January-August 2023)

PUBLICATIONS

Journal Articles

What is it for a Machine Learning Model to have a Capability?
Jacqueline Harding, Nathaniel Sharadin
British Journal for the Philosophy of Science, forthcoming

Operationalising Representation in Natural Language Processing
British Journal for the Philosophy of Science, forthcoming

What is AI Safety? What do we want it to be?
Jacqueline Harding, Cameron Domenico Kirk-Giannini
Philosophical Studies, 2025

Do As I Explain: Explanations Communicate Optimal Interventions
Lara Kirfel, Jacqueline Harding, Jeong Yeon Shin, Cindy Xin, Thomas Icard, Tobias Gerstenberg
Proceedings of the Annual Meeting of the Cognitive Science Society, 2024

AI Language Models Cannot Replace Human Research Participants
Jacqueline Harding, William D'Allesandro, N.G. Laskowski, Robert Long
AI and Society, 2023

Proxy Selection in Transitive Proxy Voting
Social Choice and Welfare, 2022

Everettian Quantum Mechanics and the Metaphysics of Modality
British Journal for the Philosophy of Science, 2021

Conference Proceedings

Toward a Formal Pragmatics of Explanation
Jacqueline Harding, Tobias Gerstenberg, Thomas Icard
Proceedings of the Annual Meeting of the Cognitive Science Society, 2025

Do As I Explain: Explanations Communicate Optimal Interventions
Lara Kirfel, Jacqueline Harding, Jeong Yeon Shin, Cindy Xin, Thomas Icard, Tobias Gerstenberg
Proceedings of the Annual Meeting of the Cognitive Science Society, 2024

Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information
Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, Willem Zuidema
Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018

Preprints

A Communication-First Account of Explanation
Jacqueline Harding, Tobias Gerstenberg, Thomas Icard
arXiv preprint, 2025

Taking AI Welfare Seriously
Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, David Chalmers
arXiv preprint, 2024

PUBLIC WRITING

Toward a More Expansive Perspective on AI Safety (Blog Post and Workshop Report)
Jacqueline Harding, César Valenzuela
HAI and McCoy Family Center for Ethics in Society, Stanford University, July 2024

AI's future worries us. So does AI's present. (Opinion Piece)
Jacqueline Harding, Cameron Domenico Kirk-Giannini
The Boston Globe, July 2023

PRESENTATIONS (+invited, *reviewed)

Goal-Directedness in AI Systems

+PAINT online lecture series, Northeastern University
(March 2025)
*MINT/Yale Workshop on Normative Philosophy of Computing, Yale Law School
(September 2024)

A New Pragmatics of Explanation

+Construction of Meaning Workshop, Stanford University
(April 2025)
+CSLI Workshop, Stanford University
(June 2024)

What is it for a Machine Learning Model to have a Capability?

+Ability Graduate Seminar, NYU
(September 2024)
+AI Benchmarking Workshop, The University of Hong Kong
(March 2024)
+Machine Intelligence and Normative Theory (MINT) Lab, Australian National University
(November 2023)

Representation in Natural Language Processing

+Philosophy of Mind Reading Group, NYU
(September 2024)
*NYU/Columbia Philosophy of Deep Learning Conference, NYU
(March 2023)

How do Language Models Track Agreement Information?

*BlackboxNLP Workshop, Empirical Methods in Natural Language Processing (EMNLP) Conference
(November 2018)

TEACHING EXPERIENCE (*primary instructor)

Stanford University

*PHIL 24H: Philosophy of Large Language Models (Autumn 2024)
PHIL 80: Mind, Matter and Meaning (Autumn 2024)
PHIL 151/251: Metalogic (Winter 2022)
PHIL 150/250: Mathematical Logic (Autumn 2021)

University of Amsterdam

Natural Language Processing 1, Artificial Intelligence MSc (Autumn 2018)
Mathematical Proof Methods for Logic, Logic MSc (grader, Autumn 2018)

University of Cambridge

1A Logic, Philosophy BA (full academic year, 2016-2017)

SELECTED AWARDS

Stanford University

Centennial Teaching Assistant Award (2022)
(awarded to ~50 Stanford TAs each year, based on faculty and student evaluations)

Human-Centered Artificial Intelligence (HAI) Graduate Fellowship (2021)
(awarded to ~12 Stanford PhD students whose work intersects with AI each year)

Patrick Suppes Fellowship in Philosophy of Science (2020-)

University of Amsterdam

Amsterdam Science Talent Scholarship (2017-2019)
(merit-based funding for MSc awarded to ~5 graduate students across the sciences)

University of Cambridge

Faculty of Philosophy

Matthew Buncombe Prize (2017) (shared)
(awarded to the best performing student on the Philosophy MPhil)

Craig Taylor Prize (2016)
(awarded to the best performing student on the Philosophy BA)

Trinity College

Travelling Studentship (2017)
(part funding for MSc)

Hyam Studentship (2016-2017)

(funding for MPhil)

Pre-Research Scholarship (2016), Senior Scholarship (2015), Junior Scholarship (2014)
(awarded on the basis of examination results in each year of the BA)

SERVICE

Departmental

Faculty of Philosophy, Stanford University

Founder and Organiser, Philosophy of AI Reading Group, 2024
Co-Head, Minorities and Philosophy (MAP) Chapter, 2022-2023
Mental Health Representative, 2021-2022

ILLC, University of Amsterdam

Master of Logic Mentor (mental health role), 2018-2019

Trinity College, Cambridge

Welfare Officer (mental health role), 2015-2016

Conference Organising

Rapporteur, Sociotechnical AI Safety Conference
Stanford University (November 2023)

Refereeing

Ethics and Information Technology, ACM FAccT Conference (x3), Mind and Language, Philosophical Studies (x3), Synthese (x2)

Last updated: June 2025