

# Jacqueline Harding

[hardingi@stanford.edu](mailto:hardingi@stanford.edu) | [jacquelineharding.github.io](https://jacquelineharding.github.io)

AOS Philosophy of AI (including AI Ethics), Philosophy of Cognitive Science, Philosophy of Science  
AOC Logic and Formal Methods, Metaphysics

## EDUCATION

### Stanford University

PhD Philosophy and Symbolic Systems (expected 2026)

### 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”)

## RESEARCH FELLOWSHIPS

### Center for AI Safety (CAIS), San Francisco

Research Fellow (January-August 2023)

## PUBLICATIONS

### Articles

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

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

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

### Comments & Replies (\*equal contribution)

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

### Under Review (draft available on request)

A paper on AI Ethics and Safety’s relationship  
Jacqueline Harding, Cameron Domenico Kirk-Giannini

A paper on machine learning models’ capabilities  
Jacqueline Harding, Nathaniel Sharadin

## **PUBLIC WRITING**

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

## **PRESENTATIONS** (+invited, \*reviewed)

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

+Machine Intelligence and Normative Theory (MINT) Lab, Australian National University,  
(November 2023)

### **Representation in Natural Language Processing**

\*NYU/Columbia Philosophy of Deep Learning Conference, New York University  
(March 2023)

### **How do Language Models Track Agreement Information?**

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

## **TEACHING ASSISTANT EXPERIENCE**

### **Stanford University**

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 with Jack Wearing  
(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**

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 (organised by Seth Lazar)  
Stanford University (November 2023)

### **Refereeing**

Philosophical Studies (x2), Ethics and Information Technology, Patterns

## **PROGRAMMING COMPETENCE**

<b>Languages</b>	Python
<b>Libraries</b>	NumPy, PyTorch, Transformers, Matplotlib