JESSICA TAYLOR

PERSONAL INFORMATION

email	jessica.liu.taylor@gmail.com
website	http://jessic.at
location	Berkeley, CA

VALUES

Truth, beauty, and life.

EDUCATION

2010–2014 Bachelor of Science, Stanford University

Stanford University

Computer Science major with concentration in Artificial Intelligence. Graduated with distinction.

2014–2015 Master of Science, Stanford University

Computer Science with concentration in Artificial Intelligence.

WORK EXPERIENCE

August 2015–June 2017 Research Fellow, MIRI

MIRI

Research at the Machine Intelligence Research Institute (MIRI) to align artificial intelligence with human values. I made progress on research problems including value learning, logical uncertainty, and decision theory.

Summer 2014 Software Engineering Intern, Google

Google

Improved inference in a machine learning project related to the Knowledge Graph and created an interface for it.

Summer 2013 Software Engineering Intern, Google

Improved some internal tools used for running experiments to improve advertising results.

Summer 2012 Software Engineering Intern, Google

Improved an App Engine test runner to add parallel test running and a better web interface, and open sourced it. https://code.google.com/p/aeta/

Summer 2011 Software Engineering Intern, Getaround

Getaround

Implemented features and fixed bugs on both web frontend and backend. Features I helped implement include SMS notifications for car rentals and improvements to the HTML rendering engine.

PUBLICATIONS

Sep. 2016 — **Logical Induction**

Not yet peer-reviewed

How might a computer algorithm assign probabilities to propositions such as "the quadrillionth digit of π is 5", far ahead of the time when their truth values can actually be computed? We present an algorithm assigning such probabilities in as asymptotically reasonable manner.

July 2016 — **Alignment for Advanced Machine Learning Systems**

Not yet peer-reviewed

As learning systems become increasingly intelligent and autonomous, what design principles can best ensure that their behavior is aligned with the interests of the operators? We present a research agenda studying this question.

June 2016 — **A Formal Solution to the Grain of Truth Problem**

Uncertainty in Artificial Intelligence

We show that reflective variants of AIXI solve a long-standing problem in game theory: how can two agents learn to model the other agent's policy in a Bayesian manner, with their beliefs having a "grain of truth" in the sense of assigning non-negligible probability to the other agent's actual policy?

Feb. 2016 — **Quantilizers: A Safer Alternative to Maximizers for Limited Optimization**

AAAI 2016 symposium

An alternative to expected utility maximization, derived using worst-case assumptions about the costs of various actions. This yields some safety properties not shared by expected utility maximization.

July 2015 — **Reflective Variants of Solomonoff Induction and AIXI**

Artificial General Intelligence

We use reflective oracles (see next paper) to define variants of Solomonoff induction and AIXI that can reason about environments containing themselves or equally powerful agents.

Oct. 2015 — **Reflective Oracles: A Foundation for Game Theory in Artificial Intelligence**

Logic, Rationality, and Interaction

When trying to define what it means for different programs to correctly predict each other's outputs, one runs into self-reference paradoxes. We use randomization to get around these, and use this result to define causal decision theory in multi-agent environments, naturally yielding Nash equilibria.

Dec. 2013 — **Learning Stochastic Inverses**

Neural Information Processing Systems

A class of algorithms for inference in Bayesian networks. It is possible to take samples from the network and use these to learn accurate conditional distributions that can be used later for inference. Co-authors: Andreas Stuhlmüller, Noah Goodman

COMPUTER SKILLS

Languages — Haskell · Python · Javascript · Java · C++ · C · C# · Scala · LaTeX · HTML · CSS

Other skills — Machine learning · Web programming (front end and back end) · Linux

OTHER INFORMATION

Interests — Artificial intelligence · Philosophy of mind · Decision theory · Ethics · Social epistemology

August 6, 2017