

AREAS OF INTEREST	AI Alignment (Machine Learning Transparency, Value Learning), Theory of Artificial Intelligence (Reinforcement Learning, Algorithmic Information Theory, Statistical Machine Learning), Economics (Agency Theory)
DEGREES	<p><i>Doctor of Philosophy in Computer Science</i>, 2016 – present University of California, Berkeley</p> <ul style="list-style-type: none">• Studying AI alignment, supervised by Stuart Russell.• Researcher at the Center for Human-Compatible AI.• GPA: 3.45/4.00 <p><i>Bachelor of Philosophy (Hons)</i>, 2012 – 2015 Australian National University</p> <ul style="list-style-type: none">• Honours in Computer Science, undergraduate studies in Mathematics and Physics.• Thesis: “Resource-bounded Complexity-based Priors for Agents”, supervised by Marcus Hutter.• GPA: 7.00/7.00, 1st Class Honours.
PUBLICATIONS	<ul style="list-style-type: none">• Loss Bounds and Time Complexity for Speed Priors. With Jan Leike and Marcus Hutter. AISTATS 2016.• Self-modification of Policy and Utility Function in Rational Agents. With Tom Everitt (lead author), Mayank Daswani, and Marcus Hutter. AGI 2016, recipient of Kurzweil Prize for Best Paper.• Exploring Hierarchy-Aware Inverse Reinforcement Learning. With Chris Cundy (lead author). GoalsRL Workshop at ICML/IJCAI/AAMAS 2018.
SELECTED AWARDS	<p><i>University Medal</i>, Australian National University 2015</p> <ul style="list-style-type: none">• Prize; awarded to students who have obtained First Class Honours (or Masters Advanced Equivalent) and demonstrated exceptional academic excellence across their studies, the highest academic prize for undergraduates. <p><i>Erin Brent Computer Science Prize</i>, Australian National University 2015</p> <ul style="list-style-type: none">• Monetary prize; awarded to the student who achieved the best Honours result in any of the degree programs relating to Computer Science, Software Engineering or Information Technology.
PROJECTS	<p><i>Structure and Representation in Neural Networks</i> Jan 1 2023 – present</p> <ul style="list-style-type: none">• Finishing my research related to cluster structure inside neural networks and representations contained in reward models, writing it up, and explaining how it forms a unified whole.
INTERNSHIPS	<p><i>Machine Intelligence Research Internship</i> 2019</p> <ul style="list-style-type: none">• Spent 3 months on research engineering team 4 days per week, while supervising a UC Berkeley intern 1 day a week.