Daniel Filan – Curriculum Vitæ

daniel_filan@berkeley.edu

AREAS OF INTEREST

AI Alignment (Value Learning, Corrigibility, Transparency), Theory of Artificial Intelligence (Reinforcement Learning, Algorithmic Information Theory, Statistical Machine Learning), Economics (Agency Theory)

DEGREES

Doctor of Philosophy in Computer Science, 2016 – present
University of California, Berkeley

- Studying AI alignment, supervised by Stuart Russell.
- GPA: 4.00/4.00

Bachelor of Philosophy (Hons), 2012 – 2015
Australian National University

- 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

- 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.

SELECTED AWARDS

University Medal, Australian National University 2015

- 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.

Erin Brent Computer Science Prize, Australian National University 2015

- 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.

National Merit Scholarship, Australian National University 2012 – 2015

- Annual funding; awarded to the top $\sim 0.5\%$ of school leavers.

Hanna Neumann Prize for Second Year Mathematics, 2013
Australian National University

- Monetary prize; awarded to the top student in second year mathematics courses.

Dean's Commendation List, Australian National University 2012

- Prize; awarded to students who achieve scores of 90 or above in all science courses in a particular year.

INTERNSHIPS

Future of Humanity Institute, Oxford University 2016

- Writing code for `agentmodels.org`, a website designed to explain the use of probabilistic programs to build models of agents and perform inference about them.

TEACHING EXPERIENCE

Teaching Assistant, MATH2322 Advanced Algebra 1 — Semester 2 2015
ANU Mathematical Sciences Institute

Teaching Assistant, MATH2320 Advanced Analysis 1 — Semester 1 2015
ANU Mathematical Sciences Institute

Teaching Assistant, COMP2610 Information Theory — Semester 2 2014
ANU Research School of Computer Science

UNDERGRAD RESEARCH

Summer Research Scholar — Summer 2013–2014
ANU Mathematical Sciences Institute

- An investigation into the theory and practice of measure-theoretic image packing.

Undergraduate Research Projects — 2013, 2014
ANU Research School of Computer Science

- Extreme state aggregation beyond MDPs: Tightness of FRL bounds.

Department of Quantum Sciences, ANU Research School of Physics and Engineering

- Proofs of impossibility theorems regarding tests of oneself being in superposition.
- An investigation into the self-gravitation of light in general relativity.