Daniel Filan

+1 (510) 646-5530 | Berkeley, CA | df@danielfilan.com | axrp.net | danielfilan.com

Work Experience

Senior Research Manager May 2025 — Present
MATS Research, Inc. *Berkeley, CA*

- Continue with previous responsibilities
- Manage a research manager

Research Manager May 2024 — May 2025
MATS Research, Inc. *Berkeley, CA*

- Conducted a mix of personal and project management for researchers entering the fields of AI alignment, security and transparency (working with the researchers themselves and their mentors)
- Ran the process of selecting mentors for future cohorts

Podcast

AI X-risk Research Podcast (AXRP) (axrp.net) Dec 2020 — Present

- Host and publish a podcast featuring long-form interviews with researchers whose work focusses on analysing and/or reducing catastrophic and existential risk from artificial intelligence
- Funded by repeat grants from the Long-Term Future Fund
- Scott Aaronson wrote of my interview with him: "The end result is … well, probably closer to my current views on this subject than anything else I've said or written!"
- Stefan Schubert tweeted that "the episodes [of AXRP about AI policy] I've listened to have been excellent and epistemically fastidious"

Education

University of California, Berkeley Berkeley, CA
Doctor of Philosophy (Computer Science) *Aug 2016 — May 2024*

- Thesis: "Structure and Representation in Neural Networks", supervised by Stuart Russell

Australian National University Canberra, Australia
Bachelor of Philosophy (Hons) *Feb 2012 — Dec 2015*

- GPA 7.0/7.0, 1st class honours, University Medal
- Primarily studied mathematics and physics
- Honours in Computer Science
- Honours thesis: "Resource-bounded Complexity-based Priors for Agents", supervised by Marcus Hutter

Publications

- **Constrained belief updates explain geometric structures in transformer representations**. Mateusz Piotrowski, Paul M. Riechers, *Daniel Filan*, Adam S. Shai. ICML, 2025.
- **Graphical clusterability and local specialization in deep neural networks**. Stephen Casper, Shlomi Hod, *Daniel Filan*, Cody Wild, Andrew Critch, Stuart Russell. PAIR^{2}Struct Workshop, ICLR, 2022.
- **Exploring hierarchy-aware inverse reinforcement learning**. Chris Cundy, *Daniel Filan*. 1st Workshop on Goal Specifications for Reinforcement Learning, FAIM, 2018.
- **Self-modification of policy and utility function in rational agents**. Tom Everitt, *Daniel Filan*, Mayank Daswani, and Marcus Hutter. AGI, 2016.
- **Loss bounds and time complexity for speed priors**. *Daniel Filan*, Jan Leike, and Marcus Hutter. AISTATS, 2016.