Michael K. Cohen

Curriculum Vitae

UC Berkeley, EECS
2121 Berkeley Way
Berkeley, CA 94704

mkcohen@berkeley.edu
www.michael-k-cohen.com
Google Scholar

RESEARCH INTERESTS

Controllable and Reliable AI; AI Security and Governance; Bayesianism

EDUCATION AND ACADEMIC APPOINTMENTS

Postdoc in Computer Science 2023-
UC Berkeley, Berkeley, CA
Supervisor: Stuart Russell

DPhil in Engineering Science 2019-2023
University of Oxford, Oxford, UK
Research Advisor: Mike Osborne

Advanced Master of Computing, *with University Medal* 2017-2019
Australian National University, Canberra, Australia
Research Advisor: Marcus Hutter

B.A. (Hons.) Chemistry, *magna cum laude* 2011-2015
Yale University, New Haven, CT
Research Advisor: Patrick Vaccaro

PUBLICATIONS

Journal articles

M. K. Cohen & M. Hutter (Forthcoming) Imitation Learning is Probably Existentially Safe. *AI Magazine.*

★ M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, & S. Russell. (2024) Regulating Advanced Artificial Agents. *Science*.

★ M. K. Cohen, M. Hutter, & N. Nanda (2022) Fully General Online Imitation Learning. *JMLR, 23(334)*.

★ M. K. Cohen, M. Hutter, & M. A. Osborne (2022) Advanced Artificial Agents Intervene in the Provision of Reward. *AI Magazine.*

M. K. Cohen, B. Vellambi, & M. Hutter (2021) Intelligence and Unambitiousness Using Algorithmic Information Theory. *IEEE Journal of Selected Areas in Information Theory.*

M. K. Cohen, E. Catt, & M. Hutter (2021) Curiosity Killed or Incapacitated the Cat and the Asymptotically Optimal Agent. *IEEE Journal of Selected Areas in Information Theory.*

D. J. Nemchick, M. K. Cohen, & P. H. Vaccaro (2016). Dual hydrogen-bonding motifs in complexes formed between tropolone and formic acid. *The Journal of Chemical Physics*, 145(20), 294-303.

Conference proceedings

B. Bucknall, S. Siddiqui, L. Thurnherr, …, M. K. Cohen, …, R. Trager (2025) In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate? In *Proc. FAccT-25*.

M. K. Cohen, M. Hutter, Y. Bengio, & S. Russell (2025) RL, But Don't Do Anything I Wouldn't Do. In *Proc. UAI-25.*

Y. Bengio, M. K. Cohen, N. Malkin, M. MacDermott, D. Fornasiere, P. Greiner, & Y. Kaddar (2025) Can a Bayesian Oracle Prevent Harm from an Agent? In *Proc. UAI-25.*

M. K. Cohen, S. Daulton, & M. Osborne (2022) Log-Linear-Time Gaussian Processes Using Binary Tree Kernels. In *Proc. NeurIPS-22*.

M. K. Cohen & M. Hutter (2020). Pessimism About Unknown Unknowns Inspires Conservatism. In *Proc. COLT-20*.

M. K. Cohen, B. Vellambi, & M. Hutter (2020). Asymptotically Unambitious Artificial General Intelligence. In *Proc. AAAI-20*.

M. K. Cohen, E. Catt, & M. Hutter (2019). A Strongly Asymptotically Optimal Agent in General Environments. In *Proc. IJCAI-19*.

Workshop proceedings

M. K. Cohen (2017). Intra-feature Random Forest Clustering. In *International Workshop on Machine Learning, Optimization, and Big Data*. 41-49. Springer, Cham.

D. Nemchick, M. K. Cohen, & P. Vaccaro (2015). Dispersion-Dominated π-Stacked Complexes Constructed on a Dynamic Scaffold. In *70th International Symposium on Molecular Spectroscopy* (Vol. 1).

Manuscripts in submission

M. K. Cohen, R. Hudson, & Y. Bengio. Superalignment Anti-Literature Review. *Submitted to ACM Computing Surveys.*

M. K. Cohen & M. Osborne. A Linear-Time, Infinite-Dimensional Extension of any Finite-Dimensional Kernel. *Submitted to JMLR.*

M. K. Cohen. Inspector Pools. *Submitted to Regulation and Governance*.

GRANTS, AWARDS, AND PRIZES

Open Philanthropy Project – Funding for a research assistant ($198,000)	2025
Open Philanthropy AI Worldviews Contest, Third Prize ($25,000)	2023
Future of Humanity Institute – DPhil Scholarship (£19,000/year + tuition)	2019-2023
Australian National University's University Medal	2019
Open Philanthropy Project – AI Scholarship ($83,530)	2017-2019

TEACHING

UC Berkeley
Co-instructor. CS188 – Artificial Intelligence. 2024
~650 undergraduates

University of Oxford
Instructor. Autonomous Intelligent Machines and Systems, AI safety module. 2020-2022
Postgraduates

UC Berkeley
Guest lecturer (x2). CS294-166 – Foundations for Beneficial AI. 2025
Postgraduates

Teach for America, Lazear Charter Academy, Oakland, CA
Computer Science Teacher 2015-2016
Middle schoolers

INVITED TALKS

Universal algorithmic intelligence reading group. Superalignment with KL regularization. September 2025.
ML Alignment and Theory Scholars. Avoiding extinction risk with pessimism. July 2025.
Center for Human-Compatible AI 2025 workshop. Guaranteed safety via pessimism. June 2025.
Simons Institute and IVADO Safety-Guaranteed LLMs workshop. *Key presenter*. Behavior of superintelligent RL agents. April 2025.
ICON Lab, UC Berkeley. Superalignment with KL Regularization. December 2024.
Mila. Superalignment with KL Regularization. November 2024.
Center for Human-Compatible AI 2024 workshop. Regulating advanced artificial agents. June 2024.
Oxford Martin School. Regulating advanced artificial agents. March 2024.
Mila. Extinction risk from RL agents and possible ways to avoid it. February 2024.
House of Commons Science and Technology Select Committee. Inquiry into the Governance of AI. January 2023.
AI for Good. Expected Behavior of Advanced Artificial Agents. November 2022.
Computational and Biological Learning Lab, University of Cambridge. Advanced Artificial Agents Intervene in the Provision of Reward. October 2022.
AI Ethics Seminar, Chalmers Institute of Technology. Advanced Artificial Agents Intervene in the Provision of Reward. October 2022.
Center for Human-Compatible AI, UC Berkeley. A Few Research Directions. September 2022.
Decision Making Group, University of Tübingen. Joint Human-AI Decision Making. March 2022.
AGI Governance Fellowship, Blavatnik School of Government. AI Existential Safety. February 2022.
Center for Human-Compatible AI Virtual Workshop, UC Berkeley. Advanced Artificial Agents Intervene in the Provision of Reward. June 2021.
Google DeepMind (Safety Team). Fully General Online Imitation Learning. February 2021.

Cambridge AI Safety Reading Group. Pessimism About Unknown Unknowns Inspires Conservatism. November 2020.
Google DeepMind (Foundations team). Pessimism About Unknown Unknowns Inspires Conservatism. September 2020.
AI Ethics London. AI Safety in Bayesian Reinforcement Learning. February 2020.
Google DeepMind (Safety team). Pessimism About Unknown Unknowns Inspires Conservatism. February 2020.
Effective Altruism Oxford. Expected Behavior of Advanced Reinforcement Learners. October 2019.
Google DeepMind (Safety team). Asymptotically Unambitious AGI. October 2019.
Center for Human-Compatible AI, UC Berkeley. Curiosity Killed the Cat and the Asymptotically Optimal Agent. September 2019.
Center for Human-Compatible AI, UC Berkeley. Asymptotically Unambitious AGI. January 2019.

ADVISING AND MENTORING

Neel Nanda, currently Team Lead at Google DeepMind
Amon Elders, currently PhD Candidate at University of Oxford
Evgenii Opryshko, currently PhD Candidate at University of Toronto
Rubi Hudson, currently PhD Candidate at University of Toronto
Jamie Bernardi, currently co-founder at AI Policy Bulletin
Alexandre Duplessis, currently master's student at University of Oxford

PAST EMPLOYMENT

Mentor, Stanford Existential Risk Initiative, Remote	2021
Visiting Researcher, Center for Human-Compatible AI, UC Berkeley, Berkeley, CA	2017-2018
Data Science Associate, Noodle.ai, Palo Alto, CA	2017
Computer Science Teacher, Lazear Charter Academy, Oakland, CA	2015-2016

SERVICE AND REVIEWING

Program Committee, 2025 Center for Human-Compatible AI Workshop	2025
Session Organizer, 2025 Center for Human-Compatible AI Workshop – AI Governance	2025

NeurIPS	2025
OECD	2024
JMLR	2020, 2021
Yale Law Journal	2025
Synthese	2021, 2022
International Journal of Production Research	2022, 2023
Journal of Consciousness Studies	2021
AAAI 2021	2020
Journal of AGI	2020
AGI 2020	2020

SELECTED MEDIA

AP News
The Conversation (co-authored with Marcus Hutter)
Southern Weekly
Southern Weekend
The Telegraph
The Times
The Independent
MIT Technology Review
CNN
NTD
Motherboard
TRT World (television)
TRT World (print)
Dubai Eye (starting at 11:10)