Michael K. Cohen

Curriculum Vitae

UC Berkeley, EECS
2121 Berkeley Way
Berkeley, CA 94704

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

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

RESEARCH INTERESTS

I research agent incentives, the theory of AI safety, and the governance of advanced agents.

REFEREED PUBLICATIONS [AR = acceptance rate]

M. K. Cohen & M. Hutter (2025) Imitation Learning is Probably Existentially Safe. ***AI Magazine*** [17% AR].

Manuscript earned a $25,000 prize.

M. K. Cohen, M. Hutter, Y. Bengio, & S. Russell (2025) RL, But Don't Do Anything I Wouldn't Do. In ***Proc. UAI-25*** [31% AR].

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*** [31% AR].

B. Bucknall, S. Siddiqui, [13 others], M. K. Cohen, [5 others], R. Trager (2025) In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate? In ***Proc. FAccT-25*** [27% AR].

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

Regulatory framework for preventing the loss of control of AI.

M. K. Cohen, S. Daulton, & M. Osborne (2022) Log-Linear-Time Gaussian Processes Using Binary Tree Kernels. In ***Proc. NeurIPS-22*** [26% AR].

M. K. Cohen, M. Hutter, & N. Nanda (2022) Fully General Online Imitation Learning. ***JMLR****, 23(334)* [20% AR].

M. K. Cohen, M. Hutter, & M. A. Osborne (2022) Advanced Artificial Agents Intervene in the Provision of Reward. ***AI Magazine*** [17% AR].

First peer-reviewed paper arguing that certain algorithms would likely seek to commandeer human infrastructure. Cited in International Scientific Report on AI Safety, Senate Testimony, and "Pause Giant AI Experiments" open letter.

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***.

M. K. Cohen & M. Hutter (2020) Pessimism About Unknown Unknowns Inspires Conservatism. In ***Proc. COLT-20*** [31% AR].

Contains explicit proof of an absence of an incentive to escape human control.

M. K. Cohen, B. Vellambi, & M. Hutter (2020) Asymptotically Unambitious Artificial General Intelligence. In ***Proc. AAAI-20*** [21% AR].

M. K. Cohen, E. Catt, & M. Hutter (2019) A Strongly Asymptotically Optimal Agent in General Environments. In ***Proc. IJCAI-19*** [18% AR].

M. K. Cohen (2017) Intra-feature Random Forest Clustering. In *International Workshop on Machine Learning, Optimization, and Big Data* [40% AR].

D. J. Nemchick, M. K. Cohen, & P. H. Vaccaro (2016) Dual hydrogen-bonding motifs in complexes formed between tropolone and formic acid. *Journal of Chemical Physics* [7% AR].

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

A. Ebtekar & M. K. Cohen. Golden handcuffs make safer AI agents. *Submitted to COLT.*

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 Journal of Mechanism and Institution Design*.

ArXiv

Y. Bengio, M. K. Cohen, D. Fornasiere, J. Ghosn, P. Greiner, M. MacDermott, S. Mindermann, A. Oberman, J. Richardson, O. Richardson, M. A. Rondeau (2025) Superintelligent agents pose catastrophic risks: Can Scientist AI offer a safer path? *arXiv preprint arXiv:2502.15657*.

GRANTS, AWARDS, AND PRIZES

Open Philanthropy Project – Funding for a research assistant ($198,000)	2025
NeurIPS – Top Reviewer Award (Free NeurIPS attendance)	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
Guest lecturer. CS188 – Artificial Intelligence. 2026
Undergraduates

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

INVITED TALKS

2026
Feb. RAND. "The No Security Principle."
Jan. Mila. "There is no inner alignment problem for supervised learning."

2025
Dec. Tsinghua. "The loss of control of AI."
Sep. Universal algorithmic intelligence reading group. "Superalignment with KL regularization."
Jul. ML Alignment and Theory Scholars. "Avoiding extinction risk with pessimism."

Jun. Center for Human-Compatible AI 2025 workshop. "Guaranteed safety via pessimism."
Apr. Simons Institute and IVADO Safety-Guaranteed LLMs workshop. **Key presenter**. "Behavior of superintelligent RL agents."

2024
Dec. ICON Lab, UC Berkeley. "Superalignment with KL Regularization."
Nov. Mila. "Superalignment with KL Regularization."
Jul. **Bill Gates residence**. "Dinner Discussion with Bill Gates." Discussion of research from me and three others.
Jun. Center for Human-Compatible AI 2024 workshop. "Regulating advanced artificial agents."
Mar. Oxford Martin School. "Regulating advanced artificial agents."
Feb. Mila. "Extinction risk from RL agents and possible ways to avoid it."

2023
Jan. **House of Commons Science and Technology Select Committee**. "Inquiry into the Governance of AI."

2022
Nov. AI for Good. "Expected Behavior of Advanced Artificial Agents."
Oct. Computational and Biological Learning Lab, University of Cambridge. "Advanced Artificial Agents Intervene in the Provision of Reward."
Oct. AI Ethics Seminar, Chalmers Institute of Technology. "Advanced Artificial Agents Intervene in the Provision of Reward."
Sep. Center for Human-Compatible AI, UC Berkeley. "A Few Research Directions."
Mar. Decision Making Group, University of Tübingen. "Joint Human-AI Decision Making."
Feb. AGI Governance Fellowship, Blavatnik School of Government. "AI Existential Safety."

2021
Jun. Center for Human-Compatible AI Virtual Workshop, UC Berkeley. "Advanced Artificial Agents Intervene in the Provision of Reward."
Feb. Google DeepMind (Safety Team). "Fully General Online Imitation Learning."

2020
Nov. Cambridge AI Safety Reading Group. "Pessimism About Unknown Unknowns Inspires Conservatism."
Sep. Google DeepMind (Foundations team). "Pessimism About Unknown Unknowns Inspires Conservatism."
Feb. AI Ethics London. "AI Safety in Bayesian Reinforcement Learning."
Feb. Google DeepMind (Safety team). "Pessimism About Unknown Unknowns Inspires Conservatism."

2019
Oct. Effective Altruism Oxford. "Expected Behavior of Advanced Reinforcement Learners."
Oct. Google DeepMind (Safety team). "Asymptotically Unambitious AGI."
Sep. Center for Human-Compatible AI, UC Berkeley. "Curiosity Killed the Cat and the Asymptotically Optimal Agent."

Jan. Center for Human-Compatible AI, UC Berkeley. "Asymptotically Unambitious AGI."

SELECTED PAST INTERNS

Evgenii Opryshko, currently PhD Candidate at University of Toronto	2025
Rubi Hudson, currently PhD Candidate at University of Toronto	2025
Alexandre Duplessis, currently master's student at University of Oxford	2025
Amon Elders, currently PhD Candidate at University of Oxford	2022
Jamie Bernardi, currently co-founder at AI Policy Bulletin	2021
Neel Nanda, currently Team Lead at Google DeepMind	2020

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, 2026
Session Organizer, 2025 Center for Human-Compatible AI Workshop – AI Governance	2025
Internship Hiring Committee, Center for Human-Compatible AI	2024
UAI (reduced load)	2026
NeurIPS (**top reviewer award**)	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

The Conversation (co-authored with Marcus Hutter)

AP News

Southern Weekly

Southern Weekend

The Telegraph
The Times
The Independent
Science
MIT Technology Review
CNN
NTD
Motherboard
TRT World (television); TRT World (print)
Dubai Eye (starting at 11:10)

REFEREES

Stuart Russell, Professor, UC Berkeley – russell@berkeley.edu

Yoshua Bengio, Professor, Université de Montréal – yoshua.bengio@mila.quebec

Michael Osborne, Professor, University of Oxford – mosb@robots.ox.ac.uk