Curriculum Vitae

Michael K. Cohen

Postdoc in Computer Science, UC Berkeley 2023-
Email: mkcohen@berkeley.edu
Supervisor: Stuart Russell

Education

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

Australian National University 2017-2019
Advanced Master of Computing Canberra, Australia
Research Advisor: Marcus Hutter
Cumulative GPA: 7.00/7.00

Yale University 2011-2015
B.A. (Hons.) Chemistry, *magna cum laude* New Haven, USA
Research Advisor: Patrick Vaccaro
Cumulative GPA: 3.90/4.00

Publications

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

Cohen, M. K., Hutter, M., Nanda, N. (2022) Fully General Online Imitation Learning. *JMLR, 23(334)*. 1-30.

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

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

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

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

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

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

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

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

Nemchick, D. J., Cohen, M. K., & Vaccaro, P. H. (2016). Dual hydrogen-bonding motifs in complexes formed between tropolone and formic acid. *The Journal of Chemical Physics*, 145(20), 204-303.

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

Awards

Australian National University's University Medal. 2019

Teaching

Co-lecturer. CS188 – Artificial Intelligence. ~700 undergraduates, UC Berkeley. 2024
Guest Lecturer/TA. Autonomous Intelligent Machines and Systems. Postgraduates, Oxford University. 2020-2022
Computer Science Teacher, Teach for America, Lazear Charter Academy (middle school), Oakland, CA. 2015-2016

Grants

Future of Humanity Institute – DPhil Scholarship (£19,000/year + tuition) 2019-2023
Open Philanthropy Project – AI Scholarship ($83,530) 2017-2019

Invited Talks

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.

Selected Media

AP News. Tech companies want to build artificial general intelligence. But who decides when AGI is attained? https://apnews.com/article/agi-artificial-general-intelligence-existential-risk-meta-openai-deepmind-science-ff5662a056d3cf3c5889a73e929e5a34

The Conversation (co-authored with Marcus Hutter). The danger of advanced artificial intelligence controlling its own feedback. https://theconversation.com/the-danger-of-advanced-artificial-intelligence-controlling-its-own-feedback-190445

Southern Weekly. Will ChatGPT bring unexpected risks to the world? https://www.toutiao.com/article/7203629400175149572/?wid=1687538101717 (ungated version)

The Telegraph. Advanced AI 'could kill everyone', warn Oxford researchers. https://www.telegraph.co.uk/news/2023/01/25/advanced-ai-could-kill-everyone-warn-oxford-researchers/

The Times. Rogue AI 'could kill everyone'. https://www.thetimes.co.uk/article/rogue-ai-could-kill-everyone-3bsfttpmv

The Independent. 'Existential catastrophe' caused by AI is likely unavoidable, DeepMind researcher warns. (headline is false) https://www.independent.co.uk/independentpremium/uk-news/artificial-intelligence-deepmind-ai-catastrophe-b2168120.html

CNN. UK parliament ponders AI's dangers. https://www.youtube.com/watch?v=RE6ThJczXtU

NTD. Michael Cohen: 'Calling for an end to giant AI experiments is not a fringe position.' (I didn't say that) https://www.ntd.com/michael-cohen-british-thought-leaders_923993.html

TRT World (Turkish Radio and Television). Future of AI: Could it really be deadly? https://www.youtube.com/watch?v=4_76qfbZ0Go

The U.S. Sun. Rewarding artificial intelligence is 'dangerous' and will affect 'survival of humanity', scientists warn. https://www.the-sun.com/tech/6515129/rewarding-artificial-intelligence-dangerous-survival-humanity-scientists-warn/

Motherboard. Google Deepmind researcher co-authors paper saying AI will eliminate humanity. (the paper did not say that) https://www.vice.com/en/article/93aqep/google-deepmind-researcher-co-authors-paper-saying-ai-will-eliminate-humanity

Dubai Eye. FIFA World Cup legacy. (starting at 11:10) https://omny.fm/shows/the-agenda-with-georgia-tolley/fifa-world-cup-legacy

TRT World. Researchers warn AI could one day 'kill everyone'. https://www.trtworld.com/magazine/researchers-warn-ai-could-one-day-kill-everyone-12777330

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, Teach for America, Lazear Charter Academy, Oakland, CA	2015-2016

Reviewing

JMLR	2020, 2021
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