Evan Hubinger

CONTACT INFORMATION
evanjhub@gmail.com
(925) 240-3826
alignmentforum.org/users/evhub
github.com/evhub

EDUCATION

Harvey Mudd College, Claremont, CA — Graduated: May 2019
B.S. in Mathematics and Computer Science — GPA: 3.912
High Distinction, Honors in Mathematics, Dean's List

The College Preparatory School, Oakland, CA — Graduated: May 2015

SUMMARY

AI safety research fellow at the Machine Intelligence Research Institute. Previously did theoretical AI safety research with Paul Christiano at OpenAI. Machine learning research experience working for HRL Laboratories and the Music Information Retrieval Lab at Harvey Mudd College. Professional software engineering experience at Google, Yelp, and Ripple. Author of the Coconut programming language.

PAPERS

An overview of 11 proposals for building safe advanced AI — May 2020
Evan Hubinger — arxiv.org/abs/2012.07532
A comparative analysis of 11 different, leading proposals for building safe advanced AI under the current machine learning paradigm. Analyzes each proposal on the four components of outer alignment, inner alignment, training competitiveness, and performance competitiveness.

Risks from Learned Optimization in Advanced Machine Learning Systems — June 2019
Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, Scott Garrabrant
arxiv.org/abs/1906.01820
Introduces the concept and the potential dangers of inner alignment and mesa-optimization, specifically discussing when a trained model might be an optimizer and how it might be misaligned.

RESEARCH EXPERIENCE

Research Fellow — November 2019 – Present
Machine Intelligence Research Institute, Berkeley, CA
• Wrote "An overview of 11 proposals for building safe advanced AI," as detailed above.
• Produced two new alternative AI safety via debate proposals, "AI Safety via Market Making" and "Synthesizing Amplification and Debate.
• Analyzed different alignment proposals from a computational complexity standpoint as in "AI safety via debate," resulting in "Alignment Proposals and Complexity Classes" and "Weak HCH Accesses EXP."
• Produced a wide variety of additional work on the Alignment Forum, including "Gradient hacking," "Chris Olah's views on AGI safety," "Understanding 'Deep Double Descent,"' "Outer alignment and imitative amplification," "Learning the prior and generalization," "Clarifying inner alignment terminology," "Homogeneity vs. heterogeneity in AI takeoff scenarios," and "Operationalizing compatibility with strategy-stealing."
• Mentored Adam Shimi, Mark Xu, and Noa Nabeshima, resulting in work such as Adam Shimi's "Universality Unwrapped" and Mark Xu's "Does SGD Produce Deceptive Alignment?"
• Was interviewed on the Future of Life Institute's AI Alignment Podcast.

Member of Technical Staff, Intern — June – September 2019
OpenAI, San Francisco, CA
• Interned on the safety team at OpenAI working under Paul Christiano to tackle theoretical safety questions on topics such as amplification and universality.
• Wrote "Relaxed adversarial training for inner alignment," an in-depth analysis of how techniques such as adversarial training or transparency tools might be used to address the inner alignment problem.
• Wrote "Are minimal circuits deceptive?" an answer to Paul Christiano's long-time open question on the malignity of minimal circuits.

Team Member — Fall 2018 – Spring 2019
HRL Laboratories Clinic Team, Harvey Mudd College, Claremont, CA
• Completed the entire design, implementation, and training of a deep reinforcement learning agent to tune quantum dots by controlling voltage gates embedded in a silicon heterostructure.

Lab Member — Spring 2018 – Spring 2019
Music Information Retrieval Lab, Harvey Mudd College, Claremont CA
• Team lead on the Sheet ID team using machine learning image retrieval techniques to identify cell phone

images of sheet music.
• Worked on using machine learning to identify beats in song recordings for the Live Song ID team.

Intern May – August 2017
Machine Intelligence Research Institute, Berkeley, CA
• Came up with the idea for, organized a research group for, and began work on the paper "Risks from Learned Optimization in Advanced Machine Learning Systems," as detailed above.
• Worked on a confidential type theory project using the Lean theorem prover.

Engineering Experience

Site Reliability Engineering Intern May – August 2017
Google, Sunnyvale, CA
• Worked as a Launch Coordination Engineer (LCE) developing the software Google uses to perform production readiness reviews of new product launches.
• Revamped the custom domain-specific language built by the LCE team to automate launch reviews.

Software Engineering Intern June – August 2016
Yelp, San Francisco, CA
• Primary author of Undebt, an open-source automated code refactoring tool with over 1,500 stars on GitHub.
• Wrote a blog post on Undebt, at the time Yelp's most popular blog post.
• Fixed errors in Yelp's configuration management that had previously taken down yelp.com.
• Rewrote Yelp's system for running large data processing operations in Elastic Map Reduce.

Software Engineering Intern June – August 2014; June – August 2015
Ripple, San Francisco, CA
• Worked on designing Interledger, a trustless system for cross-currency transactions between arbitrary agents.
• Wrote a tool to do cryptographically secure generation of wallets for financial institutions.

Personal Projects

The Coconut Programming Language October 2014 – Present
`coconut-lang.org`
• Created the Coconut programming language, a functional programming language that supports pattern-matching, algebraic data types, TRE/TCO, and compiles to any Python version.
• Coconut has over 2,300 stars on GitHub; over $1,500 in yearly support from individuals and companies including TripleByte and Kea on Open Collective; and has made the front page of r/Python, r/Programming, and twice on Hacker News.
• Presented on Coconut at PyCon 2017.
• Was interviewed on Coconut for TalkPython, Podcast.__init__, and Functional Geekery.

Minecraft Deep Learning November 2017 – February 2018
`github.com/evhub/minecraft-deep-learning`
Used Deep Q Learning to solve the task of navigating to a house in Minecraft during a snowstorm starting from a random location. Used imitation pre-training, Dueling Double DQN, and Boltzmann-Gumbel exploration.

BBopt September 2017 – Present
`github.com/evhub/bbopt`
Developed a universal black box optimization framework for tuning hyperparameters.

Open Source Contributions

Keras-RL Added support for Boltzmann-Gumbel exploration based on the paper "Boltzmann Exploration Done Right" and fixed an issue with the Normalized Advantage Functions implementation. November 2018, November 2017
Keras Fixed an issue involving invalid serialization of Keras models. November 2017
Conda Added support for advanced PEP 496 packaging features. May 2017
Typeshed Added type annotations for the `future_builtins` module. October 2016
Jupyter Added support for custom syntax highlighting. July 2016

Relevant Courses

Machine Learning Fall 2017
Neural Networks Fall 2017
Bayesian Statistics Spring 2018
Algorithms Spring 2018
Mathematical Analysis II Fall 2018
Representation Theory Fall 2018

Abstract Algebra	Fall 2017
Image Processing and Object Recognition	Fall 2016
Advanced Differential Equations and Linear Algebra	Summer 2016
Multivariable Calculus	Summer 2016
Discrete Mathematics	Spring 2016

Other Activities and Awards

Harvey Mudd Physics Department Rojansky Writing Award Winner May 2017
Awarded for the technical writing in my paper "Multiple Worlds, One Universal Wave Function."

Harvey Mudd Effective Altruism Club Leader (2017 – 2019) — AI Summer Fellows Program Attendee (2018, 2019) — Effective Altruism Global Attendee (2017, 2019, 2020) — World Wide Web Consortium Interledger Payments Community Group Member (2016) — National Forensics League Honor Society Outstanding Distinction (2015) — National Policy Debate Tournament of Champions Participant (2014, 2015)