Risks from learned optimization in advanced machine learning systems

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant

with special thanks to Paul Christiano, Eric Drexler,

Jan Leike, Rohin Shah, the MIRI agent foundations team,

and everyone else who provided feedback on earlier versions of this paper.

(Dated: May 22, 2019)

ABSTRACT.

I. INTRODUCTION

BODY.

[1] Gregory Farquhar, Tim Rocktäschel, Maximilian Igl, and Shimon Whiteson, "Treeqn and atreec: Differentiable tree-structured models for deep reinforcement learning," in *ICLR 2018* (2018).

[2] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg, "Scalable agent alignment via reward modeling: a research direction," arXiv (2018).

[3] Daniel Filan, "Bottle caps aren't optimisers," (2018).

[4] Eliezer Yudkowsky, "Optimization daemons," .

[5] Riceissa, "Optimization daemons," (2018).

[6] K. E. Drexler, "Reframing superintelligence: Comprehensive ai services as general intelligence," Technical Report #2019-1, Future of Humanity Institute, University of Oxford (2019).

[7] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation (2017).

[8] Paul Christiano, "What does the universal prior actually look like?" (2016).

[9] Alex Graves, Greg Wayne, and Ivo Danihelka, "Neural turing machines," arXiv (2014).

[10] Paul Christiano, "Open question: are minimal circuits daemon-free?" (2018).

[11] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei, "Reward learning from human preferences and demonstrations in atari," NeurIPS 2018 (2018).

[12] Kareem Amin and Satinder Singh, "Towards resolving unidentifiability in inverse reinforcement learning," arXiv (2016).

[13] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

[14] David Manheim and Scott Garrabrant, "Categorizing variants of goodhart's law," arXiv (2018).

[15] Paul Christiano, "Worst-case guarantees," (2019).

[16] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick, "Learning to reinforcement learn," CogSci (2016).

[17] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel, "Rl^2: Fast reinforcement learning via slow reinforcement learning," arXiv (2016).

[] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in ai safety," arXiv (2016).

[] Stuart Armstrong and Sören Mindermann, "Occam's razor is insufficient to infer the preferences of irrational agents," NeurIPS 2018 (2017).

[] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu, "Safety verification of deep neural networks," CAV 2017 (2016).

[] Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," CAV 2017 (2017).

[] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana, "Towards practical verification of machine learning: The case of computer vision systems," arXiv (2017).

[] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong, "Corrigibility," AAAI 2015 (2015).