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 25, 2019)

We analyze the type of learned optimization that occurs when a learned model (such as a neural network) is itself an optimizera situation we refer to as mesa-optimization. We believe that the possibility of mesa-optimization raises two important questions for the safety and transparency of advanced machine learning systems. First, under what circumstances will learned models be optimizers? Second, when a learned model is an optimizer, what will its objective be, and how can it be aligned? In this paper, we provide an in-depth analysis of these two primary questions and provide an overview of topics for future research.

CONTENTS

1. INTRODUCTION

BODY.

[1] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, Scalable agent alignment via reward modeling: a research direction, arXiv (2018).

[2] D. Filan, Bottle caps aren't optimisers (2018).

[3] G. Farquhar, T. Rocktäschel, M. Igl, and S. Whiteson, Treeqn and atreec: Differentiable tree-structured models for deep reinforcement learning, ICLR 2018 (2018).

[4] A. Srinivas, A. Jabri, P. Abbeel, S. Levine, and C. Finn, Universal planning networks, ICML 2018 (2018).

[5] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. de Freitas, Learning to learn by gradient descent by gradient descent, NIPS 2016 (2016).

[6] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, Rl^2: Fast reinforcement learning via slow reinforcement learning, arXiv (2016).

[7] E. Yudkowsky, Optimization daemons.

[8] J. Cheal, What is the opposite of meta?, ANLP Acuity Vol. 2 .

[9] E. Yudkowsky, Measuring optimization power (2008).

[10] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science **362**, 1140 (2018).

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

[12] R. Kumar and S. Garrabrant, Thoughts on human models, MIRI (2019).

[13] P. Christiano, What does the universal prior actually look like? (2016).

[14] A. Graves, G. Wayne, and I. Danihelka, Neural turing machines, arXiv (2014).

[15] G. Valle-Pérez, C. Q. Camargo, and A. A. Louis, Deep learning generalizes because the parameter-function map is biased towards simple functions, ICLR 2019 (2019).

[16] P. Christiano, Open question: are minimal circuits daemon-free? (2018).

[17] C. van Merwijk, Development of ai agents as a principal-agent problem (Forthcoming in 2019).

[18] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, Reward learning from human preferences and demonstrations in atari, NeurIPS 2018 (2018).

[19] J. Su, D. V. Vargas, and K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Transactions on Evolutionary Computation (2017).

[20] K. Amin and S. Singh, Towards resolving unidentifiability in inverse reinforcement learning, arXiv (2016).

[21] R. Pascanu, Y. Li, O. Vinyals, N. Heess, L. Buesing, S. Racanière, D. Reichert, T. Weber, D. Wierstra, and P. Battaglia, Learning model-based planning from scratch, arXiv (2017).

[22] D. Manheim and S. Garrabrant, Categorizing variants of goodhart's law, arXiv (2018).

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

[24] P. Christiano, What failure looks like (2019).

[25] N. Soares, B. Fallenstein, E. Yudkowsky, and S. Armstrong, Corrigibility, AAAI 2015 (2015).

[26] P. Christiano, Worst-case guarantees (2019).

[27] R. J. Aumann, S. Hart, and M. Perry, Games and Economic Behavior **20**, 102 (1997).

[28] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick, Learning to reinforcement learn, CogSci (2016).

[29] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, Concrete problems in ai safety, arXiv (2016).

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

[31] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, Safety verification of deep neural networks, CAV 2017 (2016).

[32] G. Katz, C. Barrett, D. Dill, K. Julian, and M. Kochenderfer, Reluplex: An efficient smt solver for verifying deep neural networks, CAV 2017 (2017).

[33] K. Pei, Y. Cao, J. Yang, and S. Jana, Towards practical verification of machine learning: The case of computer vision systems, arXiv (2017).

[34] P. Christiano, B. Shlegeris, and D. Amodei, Supervising strong learners by amplifying weak experts, arXiv (2018).

[35] G. Irving, P. Christiano, and D. Amodei, Ai safety via debate, arXiv (2018).

[36] Riceissa, Optimization daemons (2018).