Inner Optimization

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

with special thanks to Paul Christiano, Scott Garrabrant,

and the MIRI agent foundations team for originating many of the ideas discussed in this paper.

(Dated: February 12, 2019)

ABSTRACT.

I. INTRODUCTION

BODY.

[1] [2] [3] [4]

---

[1] John A. Wheeler, "Assessment of everett's "relative state" formulation of quantum theory," Reviews of Modern Physics (1957).

[2] Simon Saunders, Jonathan Barrett, Adrian Kent, and David Wallace, *Many Worlds?: Everett, Quantum Theory, & Reality* (Oxford University Press, 2010).

[3] Ray J. Solomonoff, "A preliminary report on a general theory of inductive inference," (1960).

[4] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," CoRR **abs/1710.08864** (2017), arXiv:1710.08864.