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]

---

[1] Hugh Everett, "The theory of the universal wave function," Princeton University Press (1957).