An overview of 11 proposals for building safe advanced AI

Evan Hubinger*

May 29, 2020

Abstract

TODO.

Contents

1 Introduction

TODO.[1]

References

[1] Ramana Kumar and Scott Garrabrant. Thoughts on human models. *MIRI*, 2019. URL `https://intelligence.org/2019/02/22/thoughts-on-human-models`.

*Special thanks to Kate Woolverton, Paul Christiano, Rohin Shah, Alex Turner, William Saunders, Beth Barnes, Abram Demski, Scott Garrabrant, Sam Eisenstat, and Tsvi Benson-Tilsen for providing helpful comments and feedback on this post and the talk that preceded it.
Research supported by the Machine Intelligence Research Institute (intelligence.org).