Research Paper

Annotated Bibliography

Chris Shepard

April 7, 2018

1 Research Question

The "alignment problem" in the field of general artificial intelligence (AI) is concerned with verifying that the values held by a general AI are consistent with those of humanity, and that they remain so as the AI is modified. I aim to explore some of the details of this question. In particular, I will investigate the current state of knowledge in the field, and what possible consequencecs exist for failing to address the problem adequately. My question, therefore, is "What is the alignment problem, and what happens if it remains unsolved?"

References

Bostrom, N. (2017). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

> This book is a detailed exploration of the topic of smarter-than-human general artificial intelligence. It covers many aspects of this, from describing what is meant by intelligence, to possible scenarios that may result from the development of such AIs. In particular, it explores the potential problems and dangers that may come from improperly designed AIs, and the likelihoods of those scenarios.

This book will be of direct value to my paper, as it is already an attempt to render this topic into layman's terms. I intend to primarily use this book to round out my understanding of key points, such as the problem of corrigibility. I expect to also be able to use it to better clarify information that I extract from my chosen technical papers, as this book and those papers overlap in the topics that they cover.

Kurzweil, R. (2014). Don't fear artificial intelligence. *Time*. Retrieved from https://web.archive.org/web/20180207071912/http://time.com/3641921/dont-fear-artificial-intelligence.

This article is written by a popular futurist who has been vocal about artificial intelligence and its progress into the future. The article briefly discusses some reasons for why the author is not concerned about the development of artificial intelligence. This includes a few examples of other existential risks to humanity, such as biotechnology, that have been kept relatively harmless while still providing great benefit.

This article provides a bit of a counterpoint to my narrative of artificial intelligence being a hard, complicated, dangerous problem. I intend to use it to highlight the popular perception of artificial intelligence, and contrast that with the perception of the people performing the safety research. In particular, I expect that I can make direct use of the author's counterexamples to illustrate some possible misperceptions about the incentives to making dangerous AI's.

Soares, N. (2016). The value learning problem. In *Ethics for artificial intelligence workshop.* New York: International Joint Conference on Artificial Intelligence. Retrieved from intelligence.org/files/ValueLearningProblem.pdf.

This paper is an overview of possible ways one might design artificial intelligences to learn and incorporate a model of the values of its operators. The paper briefly explains why this might be necessary, and then explains why the problem is hard. It then describes a few facets of the problem, such as corrigibility (the willingness of the AI to allow its code to be modified) and consistent ontology identification (whether or not the values of the AI today will persist through modifications made by the operators, or the

AI itself, going forward).

Since my paper is an overview of the potential problems and hurdles regarding general artificial intelligence, this paper directly touches on my topic. Furthermore, if I need to narrow my focus, I intend to dig into the topic of corrigibility in particular, which this paper introduces briefly. This paper's overview of different approaches to getting an AI to consider the values of its operators will also provide me with a better picture of the topic so that I will be better equiped to relate it in my paper.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Aaai workshops*. Austin, Texas: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. Retrieved from intelligence.org/files/Corrigibility.pdf.

This paper consists of an survey of the problem of corrigibility for AI's (the willingness or unwillingness of AI's to accept modifications from its operators, and the persistence of that willingness across future changes). The paper begins by outlining the problem itself, and then explores various aspects of the problem. These include situations where the AI may prevent itself from being shut down, or where it even forces itself to be shut down, and several more interesting variants in between.

This paper's in-depth coverage of the corrigibility problem is directly related to the specific topic I want to explore: the potential problems of designing smarter-than-human AI. It provides several specific examples of a problem, which I can use to build an illustration of the issue for my readers. The paper's focus on outlining the consequences of certain failure modes, as opposed to just highlighting the failure modes and moving on, will also help me relay an intuition of the issue to my readers.

Worley, G. G., III. (2018). Formally stating the ai alignment problem. *Map and Territory*. Retrieved from https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991.

This article lays out in specific detail many of the various aspects that go into the artificial intelligence (AI) alignment problem. It breaks the problem into its various

subproblems, names the relevant actors in the relevant fields, and explores different ways to analyze the problem itself through the various lenses of different academic disciplines such as decision theory or axiology (as in axioms, or the bottom-most set of rules that one takes as granted and builds from).

I intend to reference this paper for its concise introductions to the various people involved in the field, and their contributions. I may also reference how it divides the problem into various fields, if that seems to better fit the structure of my paper.