A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis

Anthony M. Barrett and Seth D. Baum

Journal of Experimental & Theoretical Artificial Intelligence 29(2) 397-414, 2017
This version dated 20 February 2017.
Preprint at: http://sethbaum.com/ac/2017_AI-Pathways.html

Background: Self-Improving Artificial Intelligence

This paper analyzes the risk of a catastrophe scenario involving self-improving artificial intelligence. An self-improving AI is one that makes itself smarter and more capable. In this scenario, the self-improvement is *recursive*, meaning that the improved AI makes an even more improved AI, and so on. This causes a *takeoff* of successively more intelligent AIs. The result is an *artificial superintelligence* (ASI), which is an AI that is significantly more intelligent and more capable than humans. In this scenario, the ASI gains control over the world and causes a major global catastrophe, potentially even killing everyone.

ASI-PATH: The ASI Pathways Model

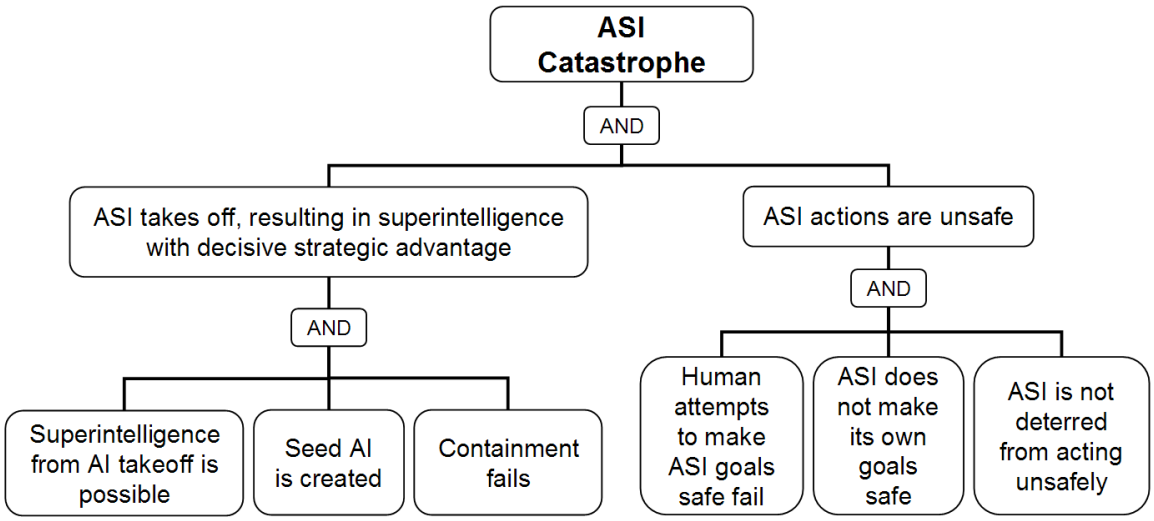
This paper introduces ASI-PATH, a model for analyzing the risk of global catastrophe from self-improving ASI. ASI-PATH is a fault tree model, which is a standard type of model in risk analysis. The model shows different pathways to ASI catastrophe. For example, the ASI could come from a specially designed AI or from an AI based on the human brain. The model connects these pathways together to show how ASI catastrophe could occur. Specific parts of the model can be assigned probabilities in order to estimate the probability of ASI catastrophe. Coming up with probabilities is a task for future research.

The Main Model Sections

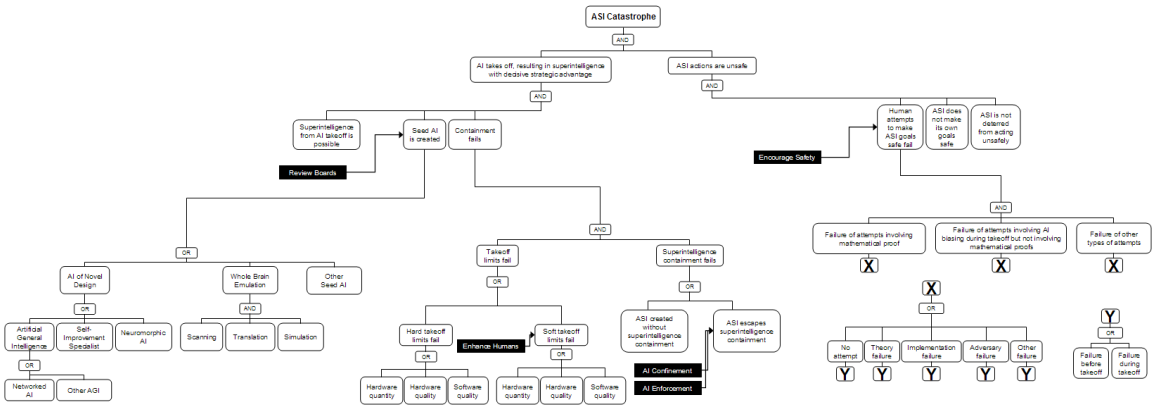
The model has two main sections. One section models how the AI gains control of the world. It has three parts: (1) the physical possibility of takeoff resulting in ASI, (2) the creation of a *seed AI* capable of recursive self-improvement, and (3) the failure of efforts to *contain* the AI so that it would not gain control of the world. The other section models how the AI uses its control to cause a catastrophe. It also has three parts: (1) the failure of human efforts to make the AI safe, (2) the AI does not make itself safe, and (3) the AI is not deterred from causing catastrophe. The two main model sections are shown in the first figure on the next page.

Risk Reduction Interventions

ASI-PATH models five interventions that could reduce the risk of ASI catastrophe: (1) Review boards to evaluate the riskiness of specific ASI projects and steer them in safer directions; (2) Encourage research into how to make ASI safer, such as by making funding available or creating a culture of safety among AI researchers; (3) Enhance human capabilities so that an AI would not gain control of the world; (4) Confine the AI so that it does not gain control of the world; and (5) Enable some AIs to prevent other AIs from gaining control of the world. Each intervention affects one or more parts of the model. If future research gets probabilities for the model parts, then the effectiveness of each intervention can be estimated.



The two main sections of the ASI-PATH model. The "AND" blocks signify that all these criteria must hold for ASI catastrophe to occur. Details are explained in the "The Main Model Sections" paragraph above.



The full ASI-PATH model as presented in this paper. Details are explained in the paper.
A high resolution version is available at: http://sethbaum.com/ac/2017_AI-Pathways2full.png