Risk Analysis and Risk Management for the Artificial Superintelligence Research and Development Process

Anthony M. Barrett and Seth D. Baum

In Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong (Eds.), 2017. *The Technological Singularity: Managing the Journey*. Berlin: Springer, pages 127-140.
Preprint at: http://sethbaum.com/ac/2017_AI-RandD.html

Background: Artificial Superintelligence

Already computers can outsmart humans in specific domains, like multiplication. But humans remain firmly in control… for now. Artificial superintelligence (ASI) is AI with intelligence that vastly exceeds humanity's across a broad range of domains. Experts increasingly believe that ASI could be built sometime in the future, could take control of the planet away from humans, and could cause a global catastrophe. Alternatively, if ASI is built safely, it may be able to solve major human problems. *This paper describes how risk analysis and risk management techniques can be used to understand the possibility of ASI catastrophe and make it less likely to happen*.

Artificial Superintelligence Risk Analysis

Risk analysis aims to understand bad events that could happen. It studies the event's probability, severity, timing and other relevant factors. For ASI risk analysis, one technique is to model the sequences of steps that could result in ASI catastrophe. Each step can then be studied to get an overall understanding of the total risk. These models are called fault trees or event trees. Creating the models and studying each step is difficult because ASI is unprecedented technology. What's happened in the past is of limited relevance. One way to study unprecedented events is to ask experts for their judgments on them. This is called expert elicitation. Experts don't always get their judgments right so it's important to ask them carefully, using established procedures from risk analysis.

Artificial Superintelligence Risk Management

There are two general ways to manage the risk of ASI catastrophe. One is to make ASI technology safer. This is a technical project that depends on the details of the specific ASI. The other is to manage the human process of ASI research and development, in order to steer it towards safer ASI and away from dangerous ASI. Risk management steps can be taken by governments, corporations, philanthropic foundations, and individual ASI researchers, among others. They can create regulations to restrict risky ASI research, covertly target risky ASI projects, and fund the development of ASI safety measures, among other things. Risk analysis and the related field of decision analysis can help people make better ASI risk management decisions. In particular, the analysis can help identify which options would be the most cost-effective, meaning that they would achieve the largest reduction in ASI risk for the amount of money spent on them.