X-Risk Analysis for AI Research

David McSharry

This paper introduces the potential and dangers of AI, and highlights the tail end risks associated with the technology. It implores the reader to think beyond the every day risks of autonomous vehicles crashing and instead think proactively to the risks we may face in the future. In particular, the paper discusses how we can make AI safer using systems safety, how we can make systems safer in the future, and reduce the risk posed by future AI. The paper then introduces some terminology from systems safety. Systems safety is explored and some principles of how we can improve the field of AI safety is discussed, such as having a safety culture around AI. Specific risks posed by AI are then discussed, such as weaponization and emergent goals. These are risks posed by the AI themselves. The paper then discusses some tractable actions we can towards making the development and deployment of AI safer, these mainly revolve around good practices for fields with potential high risks like building in safety early. Finally the paper discusses the trade off between safety and capabilities research, and recognizes that some safety research will lead to increase in capabilities so we must tread carefully, however this does not mean we should not pursue safety. For example intelligence could serve to make AI safer by better understanding our goals, but this may also lead to increased capabilities which is a downside. We should be careful in the safety-capabilties ratio of our research.
A non trivial downside of this paper is that by viewing AI safety through the lens of classical systems safety it limits the scope of what should be done to prevent AI hazards from happening. For example if we wanted to ensure that a plane would never crash again, we would ground all planes and work on safety for a long time before letting them fly, and I guess systems safety is normally not done in isolation without actually doing what you are trying to make safer over long periods of time. (I'm not sure if the counterfactual is realistic, but maybe the jist is that the paper takes for granted that capabilities increases are inevitable.)