Week 7 Paper Summaries

Joseph Camacho

August 8, 2022

X-Risk Analysis for AI Research

As the title suggests, this paper's purpose is a guide to understanding x-risks from AIs.

Lots of definitions are given. E.g., a hazard is a source of danger or potential harm. There's an important equation: Risk = Hazard x Exposure x Vulnerability (where x just means there's some relationship between them where if any of them increases, so does the risk).

Nines of reliability is introduced.

Safe design principles (like multiple layers of redundancy) are talked about.

Systemic factors can also contribute to risk (e.g. a culture of being overconfident about safety). Therefore, one way to decrease x-risks is attack these systemic factors instead of directly attacking the x-risks.

There are three main scopes of risks from a strong AI: AI System Risks (individual AI problems), Operational Risks (concerns an organization's ability to safely operate an AI in deployment), and Institutional Societal Risks (concerns the social factors that might increase risks, such as an AI arms race).

There are lots of hazards, such as enfeeblement.

Long-Term Strategies: Improve safety culture, build in safety early, improve benefit/cost ratio (right now being robust against adversarial robustness can severely decrease accuracy, for example, which is a cost companies might not want to use), prepare for crises, prioritize where to research.

When doing AI safety research, try to avoid capabilties externalities, because this can fast-forward the onset of transformative AI (which is where the x-risks come from). Note that it really is possiblle to improve safety metrics while not affecting capabilities much. E.g., PixMix does a good job of improving OOD robustness without improving accuracy in-distribution.

Judge

One strength not explicitly mentioned in the paper is that the authors included examples for many of the ideas. This greatly helps understand why the ideas they present are important, how to use them oneself, and what the ideas actually mean.