Weak 8 Readings

X-Risk Analysis for AI Research

Summary

There is currently plenty of research working to reduce current AI risks, for example self-driving car reliability and deepfake detection. However, there is much less research on preventing future risks. "Tail risks" are unlikely events in the future that could have large consequences, potentially existential consequences. In general, working on more capable AI will only increase these x-risks if there isn't more safety research done.

Risk in general can be reduced through systemic measures. Having a culture of safety is the best way to prevent disaster. In addition, better monitoring tools, more resources allocatted towards safety, smaller productivity (capability) and competitive pressures, and strong social pressures can reduce risk. With x-risk it's a little different, because 99% reliability is not good enough--one failure is the end. However, we could ideally push the nines of reliability towards infinity as long as we begin early enough.

AI safety research involves better robustness (how well does the model handle out-of-distribution queries?), monitoring (is the AI thinking and doing what it's supposed to do?), alignment (is the AI doing what humans want it to do?), and systemic safety (do the right people have access to AI?). There are plenty of issues that may arise from powerful AI, including weaponization, power-seeking AI, value lock-in, misspecified goals, and human enfeeblement. However, the sooner we seek to address these problems, the cheaper and easier they will be ("an ounce of prevention is worth a pound of cure" -BJ Franklin).

Researchers shouldn't (necessarily) stop research on more capable AI, there just needs to be a better balance with safety. Often safety and capability come together--an aligned AI might need to model human behavior very well, but doing so would require a high intelligence. Sometimes they come at the expense of one another--image detection models do moderately worse when trained against adversarial examples, but they are more robust. Instead of only improving accuracy, researchers should ensure the safety-capability ratio is not too small (one good measure could be AUROC score against accuracy).

Some current unsovled problems in AI safety include adversarial robustness (a

perfectly robust model of human values would be very important for alignment), anomaly detection (to check AI systems), transparency/honesty (does the model output its true beliefs?), power aversion (goals are currently specified so that taking power is beneficial to most AI), and moral decision-making.

Current AI systems do not constitute an x-risk. However, they are also very unsafe. If we only continue to make AI safer once we have very capable AI, we might not have enough safety measures in place when *too* capable AI is created. For this reason, some researchers prefer to work on theoretical AI safety, where an extremely capable AI is already a given.

Judgement

I thought the "safety-to-capability ratio" section was particularly intriguing. It feels too common for AI safety to be dismissed because the alternative appears to be quitting work on AI altogether. That doesn't have to be the case as long as the safety-to-capability ratio is small enough.

The statement isn't quite strong enough though. E.g. if AI is 99.999% reliable to not cause existential doom, but the safety never increases, humanity will eventually be doomed. Safety has to increase at a greater rate than capability to make x-risk less than one. It would have been nice if the paper expanded on this more precisely (i.e. mathematically), and gave more specific ways to measure risk.