4

Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)

Future highly capable AI systems (sometimes referred to as artificial general intelligence or AGI) may have a transformative effect on the world on the scale of the agricultural or industrial revolution, which could bring about unprecedented levels of global prosperity. It is by no means guaranteed however that this transformation will be a positive one without a concerted effort by the AI community to shape it that way.

As AI systems become more capable, unanticipated or unintended behavior becomes increasingly dangerous, and retrofitting safety into these more generally capable and autonomous AI systems may be difficult. Small defects in AI architecture, training, or implementation, as well as mistaken assumptions, could have a very large impact when such systems are sufficiently capable. In addition to these technical challenges, AI researchers will also confront a progressively more complex set of ethical issues during the development and deployment of these technologies.

We recommend that AI teams working to develop these systems cultivate a "safety mindset," in the conduct of research in order to identify and preempt unintended and unanticipated behaviors in their systems, and work to develop systems which are "safe by design." Furthermore, we recommend that institutions set up review boards as a resource to AI researchers and developers and to evaluate relevant projects and their progress. Finally, we recommend that the AI community encourage and promote the sharing of safety-related research and tools, and that AI researchers and developers take on the norm that future highly capable transformative AI systems "should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals." (Bostrom 2014, 254) x[xii]

Section 1 – Technical

Issue:

As AI systems become more capable, as measured by the ability to optimize more complex objective functions with greater autonomy across a wider variety of domains, unanticipated or unintended behavior becomes increasingly dangerous.

Background

Amodei et al. (2016),[xxiv] Bostrom (2014),[xxv] Yudkowsky (2008),[xxvi] and many others have discussed how an AI system with an incorrectly or imprecisely specified objective function could behave in undesirable ways. In their paper, Concrete Problems in AI Safety, Amodei et al. describe some possible failure modes, including scenarios where the system has incentives to attempt to gain control over its reward channel, scenarios where the learning process fails to be robust to distributional shift, and scenarios where the system engages in unsafe exploration (in the reinforcement learning sense). Further, Bostrom (2012)[xxvii] and Omohundro (2008)[xxviii] have argued that sufficiently capable AI systems are likely by default to adopt "convergent instrumental subgoals" such as resource-acquisition and self-preservation, unless the objective function explicitly disincentivizes these strategies. These types of problems are likely to be more severe in systems that are more capable, unless action is taken to prevent them from arising.

Candidate Recommendation

AI research teams should be prepared to put significantly more effort into AI safety research as capabilities grow. We recommend that AI systems that are intended to have their capabilities improved to the point where the above issues begin to apply should be designed to avoid those issues pre-emptively (see the next issue stated below for related recommendations). When considering problems such as these, we recommend that AI research teams cultivate a "safety mindset" (as described by Schneier [2008][xxix] in the context of computer security), and suggest that many of these problems can likely be better understood by studying adversarial examples (as discussed by Christiano [2016] [xxx]).

We also recommend that all AI research teams seek to pursue the following goals, all of which seem likely to help avert the aforementioned problems:

1. Contribute to research on concrete problems in AI safety, such as those described by Amodei et al. in *Concrete Problems in AI Safety* [xxxi] and Taylor et al. in *Alignment for Advanced Machine Learning Systems*. [xxxii] See also the work of Hadfield-Menell et al. (2016)[xxxiii] and the references therein.

2. Work to ensure that AI systems are transparent, and that their reasoning processes can be understood by human operators. This likely involves both theoretical and practical research. In particular, we recommend that AI research teams develop, share, and contribute to transparency and debugging tools that make advanced AI systems easier to understand and work with; and we recommend that AI teams perform the necessary theoretical research to understand how and why a system works at least well enough to ensure that the system will avoid the above failure modes (even in the face of rapid capability gain and/or a dramatic change in context, such as when moving from a small testing environment to a large world).

3. Work to build safe and secure environments in which potentially unsafe AI systems can be developed and tested. In particular, we recommend that AI research teams develop, share, and contribute to AI safety test environments and tools and techniques for "boxing" AI systems (see Babcock et al. [2016][xxxiv] and Yampolskiy [2012][xxxv] for preliminary work).

4. Work to ensure that AI systems fail gracefully in the face of adversarial inputs, out-of-distribution errors (see Siddiqui et al. [2016] [xxxvi] for an example), unexpected rapid capability gain, and other large context changes.

5. Ensure that AI systems are corrigible in the sense of Soares et al. (2015),[xxxvii] i.e., that the systems are amenable to shutdown and modification by the operators, and assist (or at least do not resist) the operators in shutting down and modifying the system (if such a task is non-trivial). See also the work of Armstrong and Orseau (2016).[xxxviii]

Issue:

Retrofitting safety into future more generally capable AI systems may be difficult.

Background

Different types of AI systems are likely to vary widely in how difficult they are to align with the interests of the operators. As an example, consider the case of natural selection, which developed an intelligent "artifact" (brains) by simple hill-climbing search. Brains are quite difficult to understand, and "refactoring" a brain to be trustworthy when given large amounts of resources and unchecked power would be quite an engineering feat. Similarly, AI systems developed by pure brute force might be quite difficult to align. At the other end of the spectrum, we can imagine AI systems that are perfectly rational and understandable. Realistic AI systems are likely to fall somewhere in between, and be built by a combination of human design and hill climbing (e.g., gradient descent, trial-and-error, etc.). Developing highly capable AI systems without these concerns in mind could result in systems with high levels of technical debt,[xl] leading to systems that are more vulnerable to the concerns raised in the previous issue stated above.

Candidate Recommendation

Given that some AI development methodologies will result in AI systems that are much easier to align than others, and given that it may be quite difficult to switch development methodologies late during the development of a highly capable AI system, we recommend that when AI research teams begin developing systems that are intended to eventually become highly capable, they also take great care to ensure that their development methodology will result in a system that can be easily aligned. See also the discussion of transparency tools above.

A relevant analogy for this issue is the development of the C programming language, which settled on the use of null-terminated strings [xli] instead of length-prefixed strings for reasons of memory efficiency and code elegance, thereby making the C language vulnerable to buffer overflow [xlii] attacks, which are to this day one of the most common and damaging types of software vulnerability. If the developers of C had been considering computer security (in addition to memory efficiency and code elegance), this long-lasting vulnerability could perhaps have been avoided. In light of this analogy, we recommend that AI research teams take every effort to take safety concerns into account early in the design process.

As a heuristic, when AI research teams develop potentially dangerous systems, we recommend that those systems be "safe by design," in the sense that if everything goes according to plan, then the safety precautions discussed above should not be necessary (see Christiano [2015] [xliii] for a discussion of a related concept he terms "scalable AI control"). For example, a system that has strong incentives to manipulate its operators, but which cannot due to restrictions on the system's action space, is not safe by design. Of course, we also recommend that AI research teams use all appropriate safety precautions, but safeties such as "boxes," tripwires, monitors, action limitations, and so on should be treated as fail-safes rather than as a first line of defense.

Section 2 – General Principles

Issue:

Researchers and developers will confront a progressively more complex set of ethical and technical safety issues in the development and deployment of increasingly autonomous and capable AI systems.

Background

Issues these researchers will encounter include challenges in determining whether a system will cause unintended and unanticipated harms—to themselves, system users, and the general public—as well as complex moral and ethical considerations, including even the moral weight of certain AI systems themselves or simulations they may produce (Sandberg 2014).[xliv] Moreover, researchers are always subject to cognitive biases that might lead them to have an optimistic view of the benefits, dangers, and ethical concerns involved in their research.

Candidate Recommendation

Across a wide range of research areas in science, medicine, and social science, review boards have served as a valuable tool in ensuring that researchers are able to work with security and peace of mind about the appropriateness of their research. In addition, review boards provide a valuable function in protecting institutions, companies, and individual researchers from legal liability and reputational harm.

We recommend that organizations setup review boards to support and oversee researchers working on projects that aim to create very capable and autonomous AI systems, and that AI researchers and developers working on such projects advocate that these boards be set up (see Yampolskiy and Fox [2013][xlv] for a discussion of review boards for AI projects). In fact, some organizations like Google DeepMind and Lucid AI [xlvi] have already established review boards and we encourage others to follow their example.

Review boards should be composed of impartial experts with a diversity of relevant knowledge and experience. These boards should be continually engaged with researchers from any relevant project's inception, and events during the course of the project that trigger special review should be determined ahead of time. These types of events could include the system dramatically outperforming expectations, performing rapid self-improvement, or exhibiting a failure of corrigibility. Ideally review boards would adhere to some standards or best practices developed by the industry/field as a whole, perhaps through groups like the Partnership on Artificial Intelligence.[xlvii]

Given the transformative impact these systems may have on the world, it is essential that review boards take into consideration the widest possible breadth of safety and ethical issues.

Furthermore, in light of the difficulty of finding satisfactory solutions to moral dilemmas and the sheer size of the potential moral hazard that one AI research team would face when deploying a highly capable AI system, we recommend that researchers pursue AI designs that would bring about good outcomes regardless of the moral fortitude of the research team. AI research teams should work to minimize the extent to which good outcomes from the system hinge on the virtuousness of the operators.

Issue:

Future AI systems may have the capacity to impact the world on the scale of the agricultural or industrial revolutions.

Background

The development of very capable and autonomous AI systems could completely transform not only the economy, but the global political landscape. Future AI systems could bring about unprecedented levels of global prosperity, especially given the potential impact of super intelligent AI systems (in the sense of Bostrom [2014]).[xlviii] It is by no means guaranteed that this transformation will be a positive one without a concerted effort by the AI community to shape it that way (Bostrom 2014,[xlix] Yudkowsky 2008).[xlix]

Candidate Recommendations

The academic AI community has an admirable tradition of open scientific communication. Because AI development is increasingly taking place in a commercial setting, there are incentives for that openness to diminish. We recommend that the AI community work to ensure that this tradition of openness be maintained when it comes to safety research. AI researchers should be encouraged to freely discuss AI safety problems and share best practices with their peers across institutional, industry, and national boundaries.

Furthermore, we recommend that institutions encourage AI researchers, who are concerned that their lab or team is not following global cutting-edge safety best practices, to raise this to the attention of the wider AI research community without fear of retribution. Any research group working to develop capable AI systems should understand that, if successful, their technology will be considered both extremely economically significant and also potentially significant on the global political stage. Accordingly, for non-safety research and results, the case for openness is not quite so clear-cut. It is necessary to weigh the potential risks of disclosure against the benefits of openness, as discussed by Bostrom (2016).[li] Groups like the Partnership on Artificial Intelligence[lii] might help in establishing these norms and practices.

Finally, in his book Superintelligence, philosopher Nick Bostrom proposes that we adopt a moral norm which he calls the common good principle: “Superintelligence should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals” (Bostrom 2014, 254).[liii] We encourage researchers and developers aspiring to develop these systems to take on this norm. It is imperative that the pursuit and realization of capable AI systems be done in the service of the equitable, long-term flourishing of civilization.