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

The concept of intelligence can be difficult to precisely define, and there are many proposed definitions. Legg and Hutter (2007) surveyed 70-odd definitions of intelligence, pulling out the key features and commonalities between them, and settled on the following: "intelligence measures an agent's ability to achieve goals in a wide range of environments."

In the context of autonomous and intelligent systems (A/IS), artificial general intelligence (AGI) is often used to refer to A/IS that perform comparably to humans on intellectual tasks, and artificial superintelligence (ASI or superintelligence) is commonly defined as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" (Bostrom 2014), passing some threshold of generality, well-roundedness, and versatility that present-day AI systems do not yet achieve.

Although today's state-of-the-art A/IS do not match humans in this capacity (since today's systems are only capable of performing well in limited and narrow environments or domains), many independent researchers and organizations are working on creating AGI systems (including leading AI labs like DeepMind, OpenAI, Microsoft, and Facebook's FAIR), and most AI experts expect A/IS to surpass human-level intelligence sometime this century (Grace et al. 2017).

When reasoning about the impacts that AGI systems will have, it is tempting to anthropomorphize, assume that these systems will have a "mind" similar to that of a human, and conflate intelligence with consciousness. Although it should be possible to build AGI systems that imitate the human brain, the human brain represents one point in a vast space of possible minds (Yampolskiy 2015). AGI systems will not be subject to the same constraints and engineering trade-offs as the human brain (a product of natural selection). Thus, we should not expect AGI systems to necessarily resemble human brains, just as we don't expect planes to resemble birds, even though both are flying machines. This also means that familiar faculties of intelligent entities we know like morality, compassion, and common sense will not be present by default in these new intelligences.

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

History shows that the largest drivers of change in human welfare, for better and for worse, have been developments in science, technology, and economics. Humanity's ability to drive this change is largely a function of our intelligence. Thus, one can think about building AGI as automating scientific, technological, and economic innovation. Given the disproportionate impact our intelligence has enabled our species to have on the planet and our way of life, we should expect AGI systems to have a disproportionate impact on our future, on a scale not seen since the Industrial Revolution. As such, the development of AGI systems and improvements of those systems toward superintelligence could bring about unprecedented levels of global prosperity. However, it is by no means guaranteed that the impact of these systems will be a positive one without a concerted effort by the A/IS community and other key stakeholders to align them with our interests.

As with other powerful technologies, the development and use of A/IS have always involved risk, either because of misuse or poor design (as simple examples being an assembly line worker being injured by a robotic arm or a guard robot running over a child's foot). However, as systems approach and surpass AGI, unanticipated or unintended system behavior (due to, e.g., architecture choices, training or goal specification failures, mistakes in implementation, or mistaken assumptions) will become increasingly dangerous and difficult to correct. It is likely that not all AGI-level A/IS architectures are alignable with human interests, and as such, care should be taken to analyze how different architectures will perform as they become more capable. In addition to these technical challenges, technologists will also confront a progressively more complex set of ethical issues during the development and deployment of these technologies.

In section 1 which focuses on technical issues, we recommend that A/IS 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 researchers and developers, and to evaluate relevant projects and their progress. In Section 2 which focuses on general principles, we recommend that the A/IS community encourage and promote the sharing

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

of safety-related research and tools, and that all those involved in the development and deployment take on the norm that future highly capable transformative A/IS "should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization." (Future of Life Institute 2017)

***Disclaimer:*** *While we have provided recommendations in this document, it should be understood these do not represent a position or the views of IEEE but the informed opinions of Committee members providing insights designed to provide expert directional guidance regarding A/IS. In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors or omissions, direct or otherwise, however caused, arising in any way out of the use of this work, regardless of whether such damage was foreseeable.*

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

## Section 1 — Technical

**Issue:**

**As A/IS become more capable, as measured by the ability to perform with greater autonomy across a wider variety of domains, unanticipated or unintended behavior becomes increasingly dangerous.**

### Background

A/IS with an incorrectly or imprecisely specified objective function (or goals) could behave in undesirable ways (Amodei et al. 2016, Bostrom 2014, Yudkowsky 2008). 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) and Omohundro (2008) have argued that AGI systems are likely by default to adopt "convergent instrumental subgoals" such as resource-acquisition and self-preservation, unless the system is designed to explicitly disincentivize these strategies. These types of problems are likely to be more severe in systems that are more capable (as follows from their increased optimization power and broader action space range) unless action is taken to prevent them from arising.

In order to foster safety and controllability, A/IS 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 preemptively. When considering problems such as these, teams should cultivate a "safety mindset" (as described by Schneier [2008] in the context of computer security — to anticipate and preempt adversaries at every level of design and implementation), and suggest that many of these problems can likely be better understood by studying adversarial examples (as discussed by Christiano [2016]) and other A/IS robustness and safety research threads.

Teams working on such advanced levels of A/IS should pursue the following goals, all of which seem likely to help avert the above 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, Taylor et al. in Alignment for Advanced Machine Learning Systems, and Russell et al. in Research Priorities for Robust and Beneficial Artificial Intelligence. See also the work of Hadfield-Menell et al. (2016) and the references therein.

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

2. Work to ensure that A/IS are transparent, i.e., that their internal reasoning processes can be understood by human operators. This likely involves both theoretical and practical research. In particular, teams should develop, share, and contribute to transparency and debugging tools that make the behavior of advanced A/IS easier to understand and work with; and teams should 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 infrastructure and environments for development, testing, and deployment of powerful A/IS. This work will provide some protection against risks including subversion by malicious external attackers, and unsafe behavior arising from exploratory learning algorithms. In particular, teams should develop, share, and contribute to AI safety test environments and tools and techniques for "boxing" A/IS (see Babcock et al. [2016] and Yampolskiy [2012] for preliminary work).

4. Work to ensure that A/IS "fail gracefully" (e.g., shutdown safely or go into some other known-safe mode) in the face of adversarial inputs, out-of-distribution errors (see Siddiqui et al. [2016] for an example), unexpected rapid capability gain, and other large context changes.

5. Ensure that A/IS are corrigible in the sense of Soares et al. (2015), i.e., that the systems are amenable to shutdown and modification by the operators, e.g., as with Hadfield-Menell (2017) and Russell et al. (2016), 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).

6. Explore methods for making A/IS capable of learning complex behaviors and goals from human feedback and examples, in spite of the fact that this feedback is expensive and sometimes inconsistent, e.g., as newer variants of inverse reinforcement learning attempt. See Evans et al. (2015) and Hadfield-Menell et al. (2016).

7. Build extensive knowledge layers and automated reasoning into systems to expand their contextual awareness and common sense so undesirable side effects can be determined and averted dynamically.

## Candidate Recommendations

1. Teams working on developing AGI systems should be aware that many technical robustness and safety issues are even present in today's systems and that, given more research, some corrective techniques for those can likely scale with more complex problem manifestations.

2. Teams working on developing AGI systems should be prepared to put significantly more effort into AI safety research as capabilities grow.

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

3. Teams working on developing AGI systems should cultivate a "safety mindset" like a "security mindset," vigilant of ways they can cause harm and invest in preventing those.

## Issue:

## Designing for safety may be much more difficult later in the design lifecycle rather than earlier.

### Background

Different types of AGI systems are likely to vary widely in how difficult they are to align with the interests of their operators. As an example, consider the case of natural selection, which developed an intelligent "artifact" (brains) by a process analogous to a simple hill-climbing search algorithm. Brains are quite difficult to understand, and modifying a brain to be trustworthy when given large amounts of resources and unchecked power would be extremely difficult or impossible.

Similarly, systems developed using search/optimization, especially those using multiple layers of representations, might be difficult to modify/align. At the other end of the spectrum, we can imagine systems with more principled or explicit designs that are perfectly rational, understandable, and easy to modify/align. On this spectrum, a system like AlphaGo would be closer to the search/optimization/meta end of the spectrum, and Deep Blue closer to the other.

Realistic AGI systems are likely to fall somewhere in between, and will be built by a combination of human design and search/optimization (e.g., gradient descent, trial-and-error, etc.). Developing AGI systems without these concerns in mind could result in complicated systems that are difficult or impossible to align with the interests of its operators, leading to systems that are more vulnerable to the concerns raised above.

A relevant analogy for this issue is the development of the C programming language, which settled on the use of null-terminated strings instead of length-prefixed strings for reasons of memory efficiency and code elegance, thereby making the C language vulnerable to buffer overflow 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. Paying the upfront cost in this case would have prevented much larger costs that we are still paying today. (It does require skill though to envision the types of downstream costs that can result from upstream architectural changes.)

Given that some A/IS development methodologies will result in AGI systems that are much easier to align with intentions than other methodologies, and given that it may be quite difficult to switch development methodologies and architectures late in the development of a highly capable A/IS,

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

great care should be taken by teams developing systems intended to eventually reach AGI level to ensure that their development methodology, techniques, and architecture will result in a system that can be easily aligned. (See also the discussion of transparency tools above.)

As a heuristic, when teams develop potentially dangerous systems, those systems should 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] 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 do so due to restrictions on the system's action space, is not safe by design. Of course, all appropriate safety precautions should be used, 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.

## Candidate Recommendation

When designing an advanced A/IS, researchers and developers should pay the upfront costs to ensure, to the extent possible, that their systems are "safe-by-design," and only use external restrictions on the system as fail-safes rather than as a first line of defense. This involves designing architectures using known-safe and more-safe technical paradigms as early in the lifecycle as possible.

# 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 capable A/IS.

### Background

Issues A/IS researchers and developers will encounter include challenges in determining whether a system will cause unintended and unanticipated harms — to themselves, the system's users, and the general public — as well as complex moral and ethical considerations, including even the moral weight of certain A/IS themselves or simulations they may produce (Sandberg 2014). Moreover, researchers and developers may be subject to cognitive biases that lead them to have an optimistic view of the benefits, dangers, and ethical concerns involved in their research.

Across a wide range of research areas in science, medicine, and social science, review boards have served as a valuable tool in enabling those with relevant expertise to scrutinize the ethical implications and potential risks of research activities. While A/IS researchers and developers themselves should be alert to such considerations, review boards can provide valuable additional oversight by fielding a diversity of disciplines and deliberating without direct investment in the advancement of research goals.

Organizations should set up review boards to support and oversee researchers working on projects that aim to create very capable A/IS. AI researchers and developers working on such projects should also advocate that these boards be set up (see Yampolskiy and Fox [2013] for a discussion of review boards for AI projects). There is already some precedent for this, such as Google DeepMind's ethics board (though not much is known publicly about how it functions).

Review boards should be composed of impartial experts with a diversity of relevant knowledge and experience. These boards should be continually engaged from the inception of the relevant project, 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 (international) standards or best practices developed by the industry/field as a whole, perhaps through groups like the Partnership on Artificial Intelligence, our IEEE Global Initiative, or per the Asilomar AI Principles.

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

Review boards should be complemented by other measures to draw upon diverse expertise and societal views, such as advisory groups, relevant workshops and conferences, public engagement processes, and other forums for discussion and debate. The incorporation of a wide range of viewpoints, commensurate with the breadth and scale of potential impact, will support A/IS researchers and developers in making optimal design decisions without relying solely on the oversight of review boards.

Given the transformative impact AGI 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 team would face when deploying an AGI-level system, technologists should pursue AI designs that would bring about beneficial outcomes regardless of the moral fortitude of the research team. Teams should work to minimize the extent to which beneficial outcomes from the system hinge on the virtuousness of the operators.

## Candidate Recommendation

1. Organizations working on sufficiently advanced A/IS should set up review boards to consider the implications of risk-bearing proposed experiments and development.

2. Technologists should work to minimize the extent to which beneficial outcomes from the system hinge on the virtuousness of the operators.

## Issue:

### Future A/IS may have the capacity to impact the world on a scale not seen since the Industrial Revolution.

## Background

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

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

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

Furthermore, institutions should encourage A/IS researchers and developers, who are concerned that their lab or team is not following safety best practices, to raise this to the attention of the wider A/IS community without fear of retribution. Any group working to develop capable A/IS should understand that, if successful, their technology will be considered both extremely economically and politically significant. 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) and Krakovna (2016).

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). We encourage researchers and developers aspiring to develop these systems to take on this norm. It is imperative that the pursuit and realization of AGI systems be done in the service of the equitable, long-term flourishing of civilization.

In 2017, broad coalitions of AI researchers, ethicists, engineers, businesspeople, and social scientists came together to form and to endorse the Asilomar AI Principles (Future of Life Institute 2017), which includes the relevant principles "14) Shared Benefit: AI technologies should benefit and empower as many people as possible. ... 15) Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity. ... 23) Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization."

## Candidate Recommendations

1. Adopt the stance that superintelligence should be developed only for the benefit of all of humanity.

2. De-stigmatize and remove other soft and hard barriers to AI researchers and developers working on safety, ethics, and beneficence, as well as being open regarding that work.