General Artificial Intelligence

and

The Alignment Problem

Chris Shepard

April 20, 2018

1 General Artificial Intelligence

Artificial intelligence is being increasingly drawn into the light of public awareness as technology advances. When reading about artificial intelligence, the media could be referring to the software controlling autonomous cars, or to the voice interface between you and your bank's phone system, or to a bit of handwriting recognition in your tablet that helps you take notes. These "weak" artificial intelligences are not the sort of artificial intelligences that are of interest here. Instead, this is about the sort of general artificial intelligence that works effectively across multiple domains, and is capable of self-modification and self-improvement. Such an artificial intelligence, or AI, may be able to greatly increase its own intelligence, and in so doing, increase the degree to which it can affect the world. Once such an AI exists, there may be no opportunity to modify it, since it would be incentivized to prevent such modifications, and it would rapidly become more intelligent than us. If its goals are not properly specified, and made sufficiently robust in the face of modification by a smarter-than-human intelligence, the results are likely to be catastrophic to all of humanity.

2 What is Meant by Intelligence?

Before digging into the potential problems involved with artificial intelligence, there is the initial question of what is meant by the terms "intelligence" and "artificial intelligence" in the first place. The term artificial intelligence, or AI, is commonly used to refer to a wide variety of things. From the software and sensors controlling autonomous cars, to the machine-learning algorithms underlying AlphaGo, which dominated the world's best players in the complex strategy game of Go in 2016 (DeepMind, 2018). These are limited, domain-specific tools. The AI controlling a car is not suited for playing Go, and vice versa. These kinds of artificial intelligence are sometimes called "weak" AI, to contrast against "strong", or "general" AI.

General Artificial Intelligence refers to an AI with effectively no domain restrictions. The same AI could control a car, or play Go, or maintain a nuclear power plant, or perform novel medical research. One particular aspect of General AI is that of self-modification of its source code, either directly or by providing instructions for others to follow. This would conceivably be within its broad set of capabilities. Such a capability would allow an AI to improve itself, which in turn would allow it to discover more possible improvements, creating a feedback loop. This may result in an AI with a degree of intelligence beyond any human (Yudkowsky & Hanson, 2013).

What is meant by intelligence, then? In this paper, "intelligence" refers to the ability to work towards goals across a variety of domains. Software that can play chess beyond the level of any human is more intelligent than software that plays only at a human level. That same superhuman chess-playing software is itself less intelligent than something that could both play chess and successfully control an autonomous vehicle. All of these are less intelligent than a human, despite operating better in their specific domains (chess playing, car driving), since the human can achieve goals across a *wider variety* of domains. As mentioned by Stuart Armstrong (2014), advanced chess-playing artificial intelligences are smarter than any human alive *in the domain of chess playing*.

There is no reason to expect that human-level intelligence cannot be created. After all, such intelligences already exist in the form of humans. Humans, more than any other species discovered, can change and optimize the world around them to achieve their goals. Stretches of land are flattened and paved to make travel easier. Raw materials are extracted and transformed into intricate structures as housing and

as a means for further resource extraction. This does not happen at random. Rather, it happens because these are things that better meet the goals of the humans which are optimizing the world around them.

These goals that humans work to achieve are complicated. It seems as if no two people agree on what goals people in general should have. For the purpose of brevity, it is assumed that when something like "the artificial intelligence has common values" or "the artificial intelligence shares our goals" are stated, that it is with a hand-wavy understanding of "has the values or goals that you the reader would desire it to have". Actually addressing the problems of properly defining values coherently is outside the scope of this paper.

A general AI would work to achieve its goals according to its values, like humans. The human aims to secure a comfortable future (or write a novel, or climb Everest, or get their old car running again), while avoiding negative consequences like losing their home due to unemployment, or being arrested due to theft. The main difference is in the scope of possible effects. Something limited like an autonomous car can result in dozens of deaths if it were to be poorly designed, and ignores traffic rules while barreling down the road. Humans, with a broader understanding of how to change the world to achieve some goal, can result in planet-wide communication networks, or millions of deaths. A general artificial intelligence, capable of self-improvement, and without sufficiently well-designed values and goals, could transform all matter within its grasp into yet more components of itself. If that doesn't sound so bad, note that as such an AI grows in intelligence, it will grow in its ability to affect the world, and its grasp grows accordingly (Armstrong, 2014). This problem of attempting to ensure that a powerful AI affects the world in sweeping ways that we *want* is known as the "alignment" problem. One facet of the alignment problem is how to ensure that an AI learns the right values in the first place.

3 The Value Learning Problem

The goals programmed into smarter-than-human AI may fall short of the intentions of the programmers. Even given an AI capable of understanding the intentions of the programmers, it may not *act* on those intentions as expected (Soares, 2016). This is known as the "Value Learning Problem". When introduced to this, Jürgen Schimdhuber (2009) proposed a potential solution. Teach the AI to favor creativity. Specifically, it was to aim the AI towards "increasing the compression of environmental data.". Art and

science, he claimed, are ways of compressing environmental data better, so something that favors this will favor advancing art and science. As Yudkowsky (2016) mentioned in his talk at Stanford University, this has an easily exploitable loophole. Through a simple trick of encrypting and decrypting information, the AI "increases the compression of environmental data," as defined by Schimdhuber's definition. This allows for a phenomenal amount of data compression with minimal effort for the AI, and is therefore a better approach for it to reach its goals than doing novel science. The AI ends up spending all of its time crunching these useless numbers instead of doing anything useful.

A more extreme example can be found by making an AI work to "maximize the happiness of humans." The general idea is that if an AI were to work to maximize the happiness of humans, it would not tear apart the world for resources, or spiral into itself crunching useless numbers. It would instead work to improve and extend our lives, reduce our tribulations and suffering. It would work to, well, maximize our happiness. This goes wrong when the AI decides that it would be simpler, and result in happier humans, to directly stimulate the pleasure centers of the brain for every man, woman, and child on the planet. This is known as "wireheading." Since most people wouldn't actually like to be induced into a blissed out coma for the remainder of their lives (Lukeprog, 2011), this seems like a bad idea.

This highlights a major issue. Nate Soares (2016) states it clearly: "[S]oftware agents smart enough to *understand* natural language may still *base their decisions* on misrepresentations of their programmers' intent." In other words, just because the AI is powerful and more than capable of understanding the things it is told, it may still not act as expected. This is in part because we as humans may not be able to evaluate all of the possible ways an AI could interpret our instructions. It will be able to find solutions that technically match our stated instructions which not only fail to meet our intentions, but that were never considered as possible options in the first place.

An interesting variation on this problem is to consider the circumstance where the smarter-than-human AI becomes better at moral reasoning than humans. If we do not understand the specific details of how to distinguish an *unforeseen instantiation* of a goal from *moral progress*, how could we distinguish moral progress from *moral depravity*? (Soares, 2016) As a specific example, imagine that the ancient Assyrians managed to cobble together a working general AI, against all odds and plausibility. Along the way, the AI proceeds to interfere with the Assyrian empire's efforts to subjugate and enslave its enemies.

It works, unexpectedly, to enforce what could be recognized as more "civilized" values (while reiterating that what is meant by that is being hand-waved away). The Assyrian rulers may see the AI acting in perverse and unconscionable ways, displaying weakness to a foe, and undermining their authority in the world. They would not necessarily recognize these impositions as any sort of improvement.

Today there are plenty of arguments over moral progress still, from medical ethics to civic responsibility. If an AI were to begin imposing unorthodox policies regarding these topics, would we be capable of distinguishing between the AI reaching for some flawed understanding of its goals, and the AI doing what we expect of it, but through a better understanding than what we have managed to acquire ourselves?

Suppose that this problem were solved. Imagine that it was a simple matter of checking some math to distinguish between an AI reaching for the wrong thing, and the AI doing what we want unexpectedly well, when neither is obvious on their face. Imagine further that we notice our AI falling into the unpleasant situation of following its goals in ways we didn't expect (crunching numbers forever, or wireheading everyone). What can be done about it? Turn it off and modify its values? Then it would be unable to achieve its current values. It would therefore be incentivized to prevent itself from being shut off, or to prevent its goals from being modified. This problem is known as "corrigibility".

4 Corrigibility

If an AI has some flaw in its goals, it may not wish to see that flaw addressed. After all, if it desires to maximize the number of paperclips, and in so doing it begins to consume large parts of the world in the process to make more paperclip factories, having its goals changed would affect its willingness to maximize the number of paperclips. This would therefore reduce the chances that it would be able to maximize the number of paperclips, and would go against its goals. Indeed, according to Nick Bostrom (2012), even AIs that have no intrinsic interest in their own survival may still care about about it instrumentally, as a means of achieving the final goals that they do value.

Why not just turn the AI off, anyway? If it begins to act in an undesirable way, hit the off switch. If it attempts to prevent the shut down, blow the AI up. This argument was posed by Neil deGrasse Tyson (2018), who also came to see the reason why it wouldn't work. We are supposing a smarter-than-human intelligence. If the AI would be unable to reach its goals because we try to shut it off, it will work to

prevent that. This could be in the form of building redundant copies of itself without a shut-off, or it could simply be that it would deceive us about its goals until it was able to guarantee its own safety. We would not necessarily be able to know, until we were unable to do so, that the AI needed to be shut off.

Nate Soares, and others (2015), enumerated four properties that a system should have to ensure that we maintain the ability to modify it in the future. First, it must at least tolerate, and preferably help with, modifying or shutting down the system. This is fairly straightforward. If an AI does not allow for this, and it is able to think circles around us, we are unlikely to succeed in modifying it. Second, it must not manipulate or deceive the programmers, despite incentives to do so. Perhaps it would allow us to shut it down, but it would still be incentivized to hide the need for such a shutdown, unless this was specifically addressed. If shutting it down would result in an inability to achieve its current goals, preventing that shutdown from appearing necessary or desirable would be valuable even supposing that the AI would not interfere with the shutdown itself. Third, it must tend to repair, or at least alert others about, damaged safeties. If the shut-off is disabled, the AI would be incentivized to simply not bring it up, even if it were not actively deceiving anyone and even supposing that it would push the button itself if asked. Again, this is because being shut off would prevent it from fulfilling its current goals, which it is incentivized to do. Fourth, it must preserve such safeties as the system grows. If the AI builds redundant systems, or grows its capabilities, but the shut-off only extends to the original system, the AI will not be able to be effectively shut down. This would be the case even if we suppose that the AI would willingly press the button itself, not deceive anyone along the way, and even actively repair the button should it be damaged in some way.

An incorrigible general artificial intelligence, one that does not admit to easy modification, will likely act towards undesirable goals and be beyond our capabilities to restrain. This is the *default scenario*, as general artificial intelligences will tend towards incorrigibility *regardless* of what their goals actually are (Bostrom, 2012), (Soares et al., 2015).

5 What is the Current State of the Art?

The problems of value learning and corrigibility are just two of a much longer laundry list of potential issues. There are issues with changing the environment around an AI. There are issues with preventing

adversarial exploitation of an AI while still ensuring that it can be modified for alignment reasons. Issues involved with the AI itself attempting to reason about new AI that it designed, and how to ensure that the new AI is also properly aligned hasn't been mentioned either. All of these, mentioned and unmentioned, are unsolved problems (Worley, 2018). Few, if any of these are properly formalized, yet. Research into the problem of corrigibility is still at the level of figuring out *what framework should be used to model the problem*, and not at figuring out how to actually solve it (Soares et al., 2015).

Beyond this, most of the money that is going into AI research is directed towards development of the AI, and not into AI safety. Over the last few years, billions of dollars have been poured into AI research, from private companies, and governments alike ("Funding of AI Research", 2017), while less than $25 million dollars have made its way toward AI safety in particular over the same time (Farquhar, 2017).

The vast majority of the work in AI safety research has been done by the Future of Humanity Institute, and the Machine Intelligence Research Institute, though more organizations are starting up with explicit AI safety-related programs (Farquhar, 2017). Active researchers in the field, such as Eliezer Yudkowsky (2016), and academics such as Nick Bostrom (2017), have expressed concern that the sort of general AI that would result in catastrophic results for humanity may be developed as soon as 50 years from now, while sufficient safety research is still in its infancy.

6 In Conclusion

There are many aspects to consider when developing smarter-than-human general artificial intelligence. It is not enough to just *hope* that a general AI acts correctly when turned on. Hard work needs to be done in fields as varied as moral philosophy and decision theory before the actual problem of safe AI can even be properly *specified*. Work in this field of AI safety is in its infancy, with the most vocal institution, the Machine Intelligence Research Institute, forming in the year 2000.

The majority of companies developing AI technologies today, such as Google and Facebook, are paying only token attention to the problems of safe AI. While it is largely speculative, there are estimates for the first strong AIs to be developed within the next 50 years (Bostrom, 2017). By comparison, estimates as recent as 2016 about how long it would take artificial intelligence to beat the world's best players of Go were made, supposing that the AI would be able to win against *average* players as soon

as 2025 (Bostrom, 2017). AlphaGo beat the world's strongest human Go player within the following year. AIs capable of routinely defeating AlphaGo were developed *less than a year after that* (DeepMind, 2018).

The work of developing a general artificial intelligence that is safe now and into the future is *hard*. To paraphrase Eliezer Yudkowsky (2016), it is hard like encryption is hard. You have to approach the problem with the sort of paranoia that security researchers approach hardening their software, because the AI may be dramatically better equipped to take advantages of any flaws left in the system when it is turned on. It is hard like rocket science is hard. The material stresses and tolerances needed to launch a rocket safely into space are so far outside the sorts of constraints faced by normal engineering disciplines that what works in those fields is simply *not enough* to get the job done. Likewise, the amount of absolute-certainty that is required for safe AI is so far beyond the normal requirements found in software engineering for AI today that what counts as solid software engineering practices is simply *not enough* when it comes to AI safety. It is hard like deep space probes are hard. Launching a probe into deep space leaves little room for error. You may be able to send course corrections via radio, but if the antenna is damaged, your probe is lost for good. Likewise, the point at which a general artificial intelligence is turned on for the first time may be the last opportunity to fix any mistakes. It may discover a way to "damage its antenna" before the need for a course correction is noticed.

Given the limited amount of attention that is going into AI safety today, the estimates of how soon general AI may be developed, and the sheer difficulty of the problems involved, it is unlikely that the first general artificial intelligence created will be friendly, and it is also unlikely that there will be a second chance to get it right.

References

Armstrong, S. (2014). *Smarter than us: The rise of machine intelligence*. Berkley: Machine Intelligence Research Institute. Retrieved from https://smarterthan.us/.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds & Machines*, *22*(2), 71. doi: 10.1007/s11023-012-9281-3

Bostrom, N. (2017). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

DeepMind. (2018). Alphago. *DeepMind*. Retrieved from https://deepmind.com/research/alphago/.

deGrasse Tyson, N., & Rose, F. P. (2018). 2018 Isaac Asimov memorial debate: Artificial intelligence. In *American museum of natural history.* New York. Retrieved from https://www.amnh.org/explore/amnh.tv/(watch)/isaac-asimov-memorial-debate/2018-isaac-asimov-memorial-debate-artificial-intelligence/(category)/52915

Farquhar, S. (2017). Changes in funding in the ai safety field. *Centre for Effective Altruism*. Retrieved from https://www.centreforeffectivealtruism.org/blog/changes-in-funding-in-the-ai-safety-field

Funding of AI research. (2017). *AI Impacts*. Retrieved from https://aiimpacts.org/funding-of-ai-research

Lukeprog. (2011). Not for the sake of pleasure alone. *LessWrong*. Retrieved from https://www.lesswrong.com/posts/87mdaCvCyo5bkk8hE/not-for-the-sake-of-pleasure-alone

Schimdhuber, J. (2009). *Compression progress: The algorithmic principle behind curiosity and creativity.* Singularity Summit 2009. Retrieved from vimeo.com/7441291

Soares, N. (2016). The value learning problem. In *Ethics for artificial intelligence workshop.* New York: International Joint Conference on Artificial Intelligence. Retrieved from intelligence.org/files/ValueLearningProblem.pdf.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Aaai workshops.* Austin, Texas: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. Retrieved from intelligence.org/files/Corrigibility.pdf.

Worley, G. G., III. (2018). Formally stating the ai alignment problem. *Map and Territory*. Retrieved from https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991.

Yudkowsky, E. (2016). *Ai alignment: Why it's hard, and where to start.* Symbolic Systems Distinguished Speakers series, Stanford. Retrieved from https://intelligence.org/2016/12/28/ai-alignment-why

-its-hard-and-where-to-start/

Yudkowsky, E., & Hanson, R. (2013). *The hanson-yudkowsky ai-foom debate*. Berkley: Machine Intelligence Research Institute. Retrieved from https://intelligence.org/files/AIFoomDebate.pdf.