General Artificial Intelligence

and

The Alignment Problem

Draft

Chris Shepard

April 10, 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 sort of scope-limited artificial intelligences will be referred to as "weak" artificial intelligence. For the purposes of this paper, we are interested in "strong" or "general" artificial intelligence. In particular, the sort of general artificial intelligence that 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 a "superhuman" AI exists, there may be no opportunity to modify it. If its goals are not properly specified, and made sufficiently robust in the face of modification by a smarter-than-human intelligence, the consequences can be dire.

Out of all of the possible ways to make a mind, only the tiniest portion of them result in something that shares common values. Without careful, deliberate effort, and knowledge that does not yet exist, it is likely that the first true general artificial intelligence will directly, and negatively, affect the ultimate outcome of humanity.

2 What is Meant by Intelligence?

Weak artificial intelligence, such as the AI handling speech recognition for a bank's automated phone system, or the handwriting recognition software in some tablets, share a common theme: They are limited in scope. The software that allows you to navigate your bank's interface through speech recognition is unfit for decyphering handwriting, just as the handwriting software would be a lousy choice for managing your driverless car. For the purposes of this paper, we would say that the handwriting recognition software has a "goal" of properly reading your handwriting into something machine-readable, or that the software in a driverless car has a "goal" of gettings from point A to point B while obeying a set of constraints.

A thing that has an ability to achieve its goals across a variety of domains is something that we call an "intelligent agent", or just "agent" (Legg & Hutter, 2005). 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 perhaps operating better in their specific domains (chess playing, car driving), since the human can achieve goals across a *wider variety* of domains. It depends on the purpose of the comparison, of course. As mentioned by Stuart Armstrong, advanced chess-playing artificial intelligences are smarter than any human alive *in the domain of chess playing* (2014, ch. 3).

There is no reason to expect that human-level intelligence cannot be created. After all, humans already exist. Humans, more than any other species discovered, can change the world around them to achieve their goals. They can *optimize* the world. 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, but because these are things that better meet the goals of the intelligent agents, 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, we assume that when we state something like "the artificial intelligence has common values" or "the artificial intelligence shares our goals", 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.

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 powerplant, or perform novel medical research. One particular aspect of General AI is that self-modification of its source code, either directly or by providing instructions for others to follow, 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, p. 268).

Furthermore, a general AI, like autonomous cars or humans, would work to achieve its goals according to its values. The autonomous car aims to navigate between two points while avoiding obstacles and conforming to rules like “don’t enter an intersection on a red light”. 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. The autonomous car can result in dozens of deaths if it were to be poorly designed, and ignores traffic rules while barreling down the road. The human, with a broader understanding of how to change the world to achieve some goal, can result in millions of deaths. A general artificial intelligence, capable of self-improvement, and without sufficiently well-designed values and goals, can result in (To be written...)

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”. (To be written...)

4 Corrigibility

A potential problem with increasingly intelligent artificial intelligences arises when changes need to be made to the AI. If an AI has some flaw in its goals, it may not wish to see that flaw addressed. For example, imagine that there exists an AI whose goal can be summarized as “fill this container with water”. At any given moment, if the container is not full, the AI works to fill it. (To be written...)

5 Vingean Reflection

Creating a safe general artificial intelligence, with solidly defined values that will withstand modification under the scrutiny of a smarter-than-human intelligence is a daunting challenge in itself. Assume that is somehow managed. The AI then proceeds to create a new artificial intelligence of its own. Unless further work is done, the work done to ensure the safety of the parent AI may not be sufficient for the child AI. This is what "Vingean reflection" refers to. Ensuring that intended values transfer to AIs created by other AIs. (To be written...)

6 Consequences

(To be written...)

7 Who Is Working On It?

(To be written...)

8 In Conclusion

There are many aspects to consider when developing smarter-than-human general artificial intelligence. It is not enough to just *hope* that we get it right. 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. (needs citation) By comparison, estimates about how long it would take artificial intelligence to beat the world's best players of Go were targetted as late as 2030, but AlphaGo beat the world's strongest human Go player within the following year. (needs citation, Melhauser quote?)

The work of developing a general artificial intelligence that is safe now and into the future is *hard*. 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 time we have to fix any mistakes. It may discover a way to "damage its antenna" before we realize that a course correction is needed. (Needs Yudkowsky Stanford talk citation) 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 shear difficulty of the problems involved, it is unlikely that the first general artificial intelligences 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/.

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

Legg, S., & Hutter, M. (2005). A universal measure of intelligence for artificial agents. In *Ijcai-05.* Edinburgh, Scotland, UK: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence. Retrieved from http://www.ijcai.org/Proceedings/05/Papers/post-0042.pdf.

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.

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