Technical and Ethical Considerations

for Artificial General Intelligence

Shaurya Singh

IM-UH 3313 Robota Psyche

Final Paper

Professor Michael Shiloh

New York University Abu Dhabi

Spring 2021

Abstract

Artificial General Intelligence (AGI) as a subject, especially against the implied backdrop of a robot-apocalypse gains a lot of traction in popular culture. However, this paper takes a look at the evidence - how far along are we, and what would be some technical hurdles that will need to be overcome before we can think about a true AGI application - spoiler alert, it is a long way down the line and probably not within our lifetimes. Most, if not all, advancements that we have seen in AI is in the domain we call "narrow AI", which refers to solving finite sample space deterministic problems using a large amount of data. There is not much out there that accurately simulates what one may call "creative thinking" - and creative thinking is often only the intermediary step towards the final conclusive state of AGI we envision - that is, sentient AI. It is not clear how many years it will take before making the jump from modern narrow AI to systems that can creatively think, and the next jump from there on to sentience is one for which we don't even have a vague guess. However, that is not to say that an analysis of the modern landscape is fruitless, or I'd rather hope not - this paper is predominantly based on the assumption that this analysis is a worthwhile activity to undertake.

Introduction

"The first ultraintelligent machine is the last invention that man need ever make."

I.J. Good (1965)

I.J. Good may not be a household name in 2021, but as a close associate of Alan Turing during the second World War, the British mathematician and cryptologist was uniquely positioned to look ahead into the future back into the early days of humanity's dreams for replicating human-like intelligence.

A fundamental problem with discerning the current state of AGI development is that the sharpest minds of our time do not help our cause. Whereas Elon Musk[1] has often warned over various podcast appearances (Joe Rogan Experience as well as the Lex Fridman[2] podcast) that AGI is inevitable and coming in the next 10-25 years, some others including Jerome Pesenti[3] and Yann LeCun have taken strong anti-AGI positions in that we are nowhere close to mimicking human intelligence.

Before one begins to delve into any form of meaningful discussion about the current state of artificial intelligence development, it is important to break down a few definitions. I have tried to do that below, often having to combine definitions from multiple sources in order to arrive at the scopes that I think are an accurate approximation of reality, both from a technical standpoint as well as from a public opinion one.

Narrow AI: Narrow AI refers to using large datasets to solve problems that have a set number of outcomes. These are often called bounded problems, where the choices at any turn are finite and deterministic. IBM's Deep Blue, a chess-playing software that famously beat Gary Kasparov, is a widely known example of this subcategory. Deepmind's AlphaGo is another recent example - it is also a significant one because while Go is also a finite bounded game like chess, the sample space itself dwarfs that of chess and therefore has provided us with real evidence that the sample spaces over which narrow AI can be effective are continuously expanding over time as both software development and processing power improves over time.

Artificial General Intelligence : AGI is arguably a very human-centric term. Whereas the exact nuances of this term may vary, it is most commonly used to denote an artificial intelligence system that has fluid intelligence similar to that of a human being. This includes solving unbound problems (such as a business strategy decision) and human-like creativity (whether it be painting or writing code). One school of thought proposes that a strong NLP-core is a prerequisite for AGI in order for it to sustain human-like conversations. This may or may not be true in the long run, but in any case would not be a limiting factor towards the advancement of AGI given that NLP has actually made some of the largest strides in competency over the last decade whereas convincing simulations of creative thought border on the impossible at the moment. The ability to reason generally across a wide range of open domain questions, as well as a superhuman ability to self-improve over time are also considered some of the fundamental characteristics of AGI.

The current state of AGI

I am inclined to agree with Pesenti and LeCun in the assertion that AGI is not anywhere near the range of our technical abilities at the moment - however, this does that mean that progress has not been made. If one were to view general intelligence as a sum total of a finite (although extraordinarily large) number of bounded domains - where:

General Intelligence = sum(chess, medical decision-making, business strategy, artistic creativity...) ,

then there has been undeniable progress towards building intelligent software that could, one day, be versatile enough to simulate general intelligence. The fundamental problem is the number of terms in the right-hand-side summation, which is why the ability to self-improve is seen as critical to the development of AGI because only via automated pushes in versatility could we hope to conquer a meaningful number of domains.

In other words, true AGI would not be significantly different in its design compared to a narrow system like AlphaGo - however, it would have the ability to perform to the same level in multiple domains without a need for excessive re-programming. In fact, I would argue that the level of competency in *each* domain need not be as high as AlphaGo's competency in Go - an AGI application with above average human intellect at scale over a variety of domains would be a desirable middle ground from a technical perspective instead of trying to replicate AlphaGo's world-beating standards across thousands of domains. In addition, there is some evidence to suggest that average human intellect is somewhat quantifiable through testing - however this will require adaptation to steer away from measuring things that may be meaningless in AI terms - such as memorization. (Legg, Hutter et al, 2007 - DeepMind research)

Of all the different projects being run across corporations and universities, the one with the widest array of applications as well as successful commercial precedent is IBM's Watson. Watson is currently used for a variety of applications - utilization management decisions in hospitals and medical research, content moderation at scale for networks like Twitter[4], data-driven decision making for advertisement technology systems, as well as weather forecasting.

Watson represents the state-of-the-art for several AI sub-fields: automated reasoning, open domain question answering, as well as natural language processing. While Watson is not the leading system for niche problems like chess or Go, its appeal lies in its ability to work reasonably well across multiple bounded domains when supplied with enough data. This is not to say that it has succeeded completely - while earlier versions of medical applications envisioned Watson as an AI doctor, the truth is that it is closer to an AI medical admin assistant at the moment. However, the versatility (even with limited competency) is an extremely encouraging development given that it is a closer approximation of human intelligence than the narrow AI applications built by Deepmind and the like, which while exceptional in their own right, are useful only in the context of artificial game environments with finite rules. (Fjelland, 2020)

Technical Challenges towards true AGI

The fundamental challenge towards replicating human intelligence is that our understanding of it is rather limited. This is true from both a philosophical as well as a biological perspective.

Defining human intelligence is exceedingly difficult in view of the range of abilities and competency across people. There is no way to map the quantitative difference in ability between me and Mozart, say, and even if there was, that framework would fail when asked to map the same difference between me and Magnus Carlsen simply because the exact variable measured is so different. It is, therefore, difficult to identify the goals for AGI - what level of competency must AGI development strive for in each domain, and if for argument's sake one were to settle on the "average" level of intelligence, then how do we go about measuring that the average musical ability in the human population is, for example?

One philosophical resolution of this question lies in one of the fundamental prerequisites we defined in the Introduction: AGI must have a superhuman ability for self-improvement. This assertion posits that it does not matter what the starting point of AGI competency is, because given enough time AGI systems will gain superhuman competency in every domain they're deployed in. However, this resolution is exactly that - philosophical, because it does not offer us any ideas on how this self-improvement mechanism may be implemented, and in my mind versatile automated self-improvement is one of the major challenges that will need to be addressed before we can think about achieving true AGI.

Our biological understanding of the origin of human intelligence is also quite limited. To begin with, all our brain-mimicking abstractions such as neural networks take a neuron as a discrete unit of intelligent-hardware. This works for simpler narrow problems and is also convenient, because of the binary nature of silicon-based systems which can mimic firing and non-firing neurons in terms of 0s and 1s. However, the natural reality is that the neuron is far from a discrete unit in itself - each neuron is a complex structure with a full-fledged cell mechanism inside it, and our understanding of how these sub-units work together to form the neural outputs that we observe is very limited. (Ravikant, 2019)

As a proponent of AGI one must believe in the purely physiological nature of consciousness and intelligence, because reconciling spiritual/supernatural origins for these human traits with AGI research is quite difficult. However, even if one agrees that human intelligence is a result of a complicated physiological system, the truth is that we do not know exactly how the brain functions at the fundamental most discrete-level, and our attempts at mimicking its functionality have attempted to create a high-level copy that abstracts away the internal workings of a neuron into a binary 0 and 1 output. Furthermore, the brain is not a uniformly composed of the same neural structure everywhere: the pre-frontal cortex for instance is different in its functioning than the hippocampus or the brain stem - and these differences have not been given enough attention in contemporary AGI research so far.

The ethical arguments for and against AGI

Meredith Broussard, who is a Professor of data journalism at NYU, argues in her book "*Artificial Unintelligence*" that the fixation with solving every problem with technology has led to suboptimal solutions in various fields - but especially related to fields which impact socioeconomic equity. (Broussard, 2018).

It is easy to see that the advent of AGI applications would only accelerate the trend of using technology to make decisions, and Broussad's arguments about how tainted datasets that are skewed due to the biases of the previously human decision-makers may also cause AGI systems to make biased decisions is not very far-fetched, especially in the context of this paper.

A case study of IBM's ALERT system developed for Kansas City is often cited to support the idea that AI-enabled decision-making systems might not always work as intended. Using datasets marred with decades of systemic racism that saw a disproportionate number of black people convicted of more serious charges, ALERT was susceptible to the same biases with its predictive models - ironically the one problem it was meant to solve. IBM has often maintained that the system worked as intended (IBM Archives), but as Charlton Mcilwain, a Professor of Media, Culture, and Communications at NYU writes, this was decisively not the case as the predictive systems fell prey to biased datasets that furthered racial inequity. (Mcilwain, 2020)

This is not to say that Broussard's argument is watertight. One argument given against AGI susceptibility towards existing biases would be that unlike traditional narrow applications, true AGI would be able to discern past bias and apply corrective measures to predictions. However, this is a philosophical solution that creates more problems than it addresses - because the "correct" way to undo past biases is subject to much socio-political debate and culture wars between the right and the left. It is not unreasonable to expect, therefore, that multiple variations of AGI applications may be born which reflect the leanings of their makers - simply because there is no "correct" solution to many open ended social problems. In such a scenario, Broussard suggests that we further the cause of "algorithmic accountability", wherein even though politically-charged development of AGI applications is not outlawed, firms are under obligation to disclose how their AGI applications make decisions and what political frameworks they encompass. This could hypothetically work as governments could then have a choice of AGI applications to help with governance based on their own political philosophies - despite the obvious downfalls of radical bias in any system (AGI or human), there is merit in the automation AGI systems could provide for governmental tasks.

Conclusion

In conclusion, the technical challenges that stand in the way of AGI development can be summed up as follows:

1. It is difficult to define what human intelligence looks like as a sum of specific domain bounded abilities - meaning the goals for true AGI are vague.
2. The ability to self-learn at scale, with superhuman speed, without the need for extensive repurposing across domains, will be key towards expanding the scope of narrow algorithmic AI applications towards something resembling AGI. How this can be achieved is unclear.
3. Current software abstracts away the nuances of how biological brains work - but the rising consensus amongst researchers is that this is important and that the advancement of AGI will be closely related to advancements in neuroscience.

Even if these seemingly insurmountable challenges are subdued, the sociopolitical debates around AGI will continue to rage for a while. The most likely version of the future seems to be that the merits of AGI would lie in its automation strengths and not in its ability to solve socio-economic problems in an absolute manner. Of course, there is no telling what point #2, if accomplished, would lead us to - of all the challenges discussed in the paper, it remains by far the most daunting as well as the one with the highest unexpected payoffs in the long run.

Notes

1. Musk is by no means a technical expert on AI, but his position as one of DeepMind's principal investors as well as a founder of OpenAI would lend him some informed opinions, one would imagine.
2. Fridman is an AI-researcher at MIT and seems to agree with some of Musk's takes.
3. Pesenti is the Head of AI at Facebook, while LeCun is the Chief AI Scientist at Facebook and a Professor of Mathematical Sciences at NYU.
4. While this may seem impressive, in recent years the problems faced by Twitter and Facebook have made it very clear that modern AI, including Watson, is not very good at content moderation at scale.

References

1. Fjelland, Ragnar. “Why General Artificial Intelligence Will Not Be Realized.” *Nature News*, Nature Publishing Group, 17 June 2020, www.nature.com/articles/s41599-020-0494-4.
2. “CEO Naval Ravikat Says AI Fears Are Overblown | Joe Rogan.” *YouTube*, YouTube, 4 June 2019, www.youtube.com/watch?v=QGYbbLWn-IE.
3. Legg, Shane, and Marcus Hutter. “Universal Intelligence: A Definition of Machine Intelligence.” *Minds and Machines*, Springer Netherlands, 10 Nov. 2007, link.springer.com/article/10.1007%2Fs11023-007-9079-x.
4. Broussard, Meredith. “Artificial Unintelligence: How Computers Misunderstand the World.” MIT Press, 2018 (Hardcover). ISBN: 9780262038003.
5. “Catching the Bad Guys.” *IBM Archives: Catching up the Bad Guys*, www.ibm.com/ibm/history/exhibits/valueone/valueone_bad.html.
6. Mcilwain, Charlton. “Critical Figures: Charting the History of 'Black Software' in Tech.” *The Reboot*, 16 Dec. 2020. thereboot.com/critical-figures-charting-the-history-of-black-software-in-tech/.