Four Background Claims

July 24, 2015 | Nate Soares (https://intelligence.org/author/nate/) | Analysis (https://intelligence.org/category/analysis/)

MIRI's mission is to ensure that the creation of smarter-than-human artificial intelligence has a positive impact. Why is this mission important, and why do we think that there's work we can do today to help ensure any such thing?

In this post and my next one, I'll try to answer those questions. This post will lay out what I see as the four most important premises underlying our mission. Related posts include Eliezer Yudkowsky's "Five Theses (https://intelligence.org/2013/05/05/five-theses-two-lemmas-and-a-couple-of-strategic-implications/)" and Luke Muehlhauser's "Why MIRI? (https://intelligence.org/2014/04/20/why-miri/)"; this is my attempt to make explicit the claims that are in the background whenever I assert that our mission is of critical importance.

Claim #1: Humans have a very general ability to solve problems and achieve goals across diverse domains.

We call this ability "intelligence," or "general intelligence." This isn't a formal definition (https://intelligence.org/2013/06/19/what-is-intelligence-2/) — if we knew *exactly* what general intelligence was, we'd be better able to program it into a computer — but we do think that there's a real phenomenon of general intelligence that we cannot yet replicate in code.

Alternative view: There is no such thing as general intelligence. Instead, humans have a collection of disparate special-purpose modules. Computers will keep getting better at narrowly defined tasks such as chess or driving, but at no point will they acquire "generality" and become significantly more useful, because there is no generality to acquire. (Robin Hanson (http://www.overcomingbias.com/2014/07/limits-on-generality.html) has argued for versions of this position.)

Short response: I find the "disparate modules" hypothesis implausible in light of how readily humans can gain mastery in domains that are utterly foreign to our ancestors. That's not to say that general intelligence is some irreducible occult property; it presumably comprises a number of different cognitive faculties and the interactions between them. The whole, however, has the effect of making humans much more cognitively versatile and adaptable than (say) chimpanzees.

Why this claim matters: Humans have achieved a dominant position over other species not by being stronger or more agile, but by being more intelligent. If some key part of this general intelligence was able to evolve in the few million years since our common ancestor with chimpanzees lived, this suggests there may exist a relatively short list of key insights that would allow human engineers to build powerful generally intelligent AI systems.

Further reading: Salamon et al., “How Intelligible is Intelligence? (https://intelligence.org/files/HowIntelligible.pdf)”

Claim #2: AI systems could become much more intelligent than humans.

Researchers at MIRI tend to lack strong beliefs about *when* smarter-than-human machine intelligence will be developed. We do, however, expect that (a) human-equivalent machine intelligence will eventually be developed (likely within a century, barring catastrophe); and (b) machines can become significantly more intelligent than any human.

Alternative view #1: Brains do something special that cannot be replicated on a computer.

Short response: Brains are physical systems, and if certain versions of the Church-Turing thesis (https://en.wikipedia.org/wiki/Church%E2%80%93Turing_thesis) hold then computers can in principle replicate the functional input/output behavior of any physical system. Also, note that “intelligence” (as I’m using the term) is about problem-solving capabilities: even if there were some special human feature (such as qualia (http://www.iep.utm.edu/hard-con/)) that computers couldn’t replicate, this would be irrelevant unless it prevented us from designing problem-solving machines.

Alternative view #2: The algorithms at the root of general intelligence are so complex and indecipherable that human beings will not be able to program any such thing for many centuries.

Short response: This seems implausible in light of evolutionary evidence. The genus *Homo* diverged from other genera only 2.8 million years ago, and the intervening time — a blink in the eye of natural selection — was sufficient for generating the cognitive advantages seen in humans. This strongly implies that whatever sets humans apart from less intelligent species is not extremely complicated: the building blocks of general intelligence must have been present in chimpanzees.

In fact, the relatively intelligent behavior of dolphins suggests that the building blocks were probably there even as far back as the mouse-sized common ancestor of humans and dolphins. One could argue that mouse-level intelligence will take many centuries to replicate, but this is a more difficult claim to swallow, given rapid advances (https://www.youtube.com/watch?v=GYQrNfSmQ0M) in the field of AI. In light of evolutionary evidence and the last few decades of AI research, it looks to me like intelligence is something we will be able to comprehend and program into machines.

Alternative view #3: Humans are already at or near peak physically possible intelligence. Thus, although we may be able to build human-equivalent intelligent machines, we won’t be able to build superintelligent machines.

Short response: It would be surprising if humans were perfectly designed reasoners, for the same reason it would be surprising if airplanes couldn't fly faster than birds. Simple physical calculations bear this intuition out: for example, it seems well possible, within the boundaries of physics, to run a computer simulation of a human brain at thousands of times the normal speed.

Some expect that speed wouldn't matter, because the real bottleneck is waiting for data to come in from physical experiments. This seems unlikely to me. There are many interesting physical experiments that can be sped up, and I have a hard time believing that a team of humans running at a 1000x speedup would fail to outperform their normal-speed counterparts (not least because they could rapidly develop new tools and technology to assist them).

I furthermore expect it's possible to build *better* reasoners (rather than just *faster* reasoners) that use computing resources more effectively than humans do, even running at the same speed.

Why this claim matters: Human-designed machines often knock the socks off of biological creatures when it comes to performing tasks we care about: automobiles cannot heal or reproduce, but they sure can carry humans a lot farther and faster than a horse. If we can build intelligent machines specifically designed to solve the world's largest problems through scientific and technological innovation, then they could improve the world at an unprecedented pace. In other words, AI matters.

Further reading: Chalmers, "The Singularity: A Philosophical Analysis (http://consc.net/papers/singularity.pdf)"

Claim #3: If we create highly intelligent AI systems, their decisions will shape the future.

Humans use their intelligence to create tools and plans and technology that allow them to shape their environments to their will (and fill them with refrigerators, and cars, and cities). We expect that systems which are even more intelligent would have even more ability to shape their surroundings, and thus, smarter-than-human AI systems could wind up with significantly more control over the future than humans have.

Alternative view: An AI system would never be able to out-compete humanity as a whole, no matter how intelligent it became. Our environment is simply too competitive; machines would have to work with us and integrate into our economy.

Short response: I have no doubt that an autonomous AI system attempting to accomplish simple tasks would initially have strong incentives to integrate with our economy: if you build an AI system that collects stamps for you, it will likely start by acquiring money to purchase stamps. But what if the system accrues a strong technological or strategic advantage?

As an extreme example, we can imagine the system developing nanomachines and using them to convert as much matter as it can into stamps; it wouldn't necessarily care whether that matter came from "dirt" or "money" or "people." Selfish actors only have an incentive to participate in the economy when their gains from trade are greater than the net gains they would get by ignoring the economy and just taking the resources for their own.

So the question is whether it will be possible for an AI system to gain a decisive technological or strategic advantage. I see this as the most uncertain claim out of the ones I've listed here. However, I expect that the answer is still a clear "yes."

Historically, conflicts between humans have often ended with the technologically superior group dominating its rival. At present, there are a number of technological and social innovations that seem possible but have not yet been developed. Humans coordinate slowly and poorly, compared to what distributed software systems could achieve. All of this suggests that if we build a machine that does science faster or better than we can, it could quickly gain a technological and/or strategic advantage over humanity for itself or for its operators. This is particularly true if its intellectual advantage allows it to socially manipulate humans, acquire new hardware (legally or otherwise), produce better hardware, create copies of itself, or improve its own software. For good or ill, much of the future is likely to be determined by superintelligent decision-making machines.

Why this claim matters: Because the future matters. If we want things to be better in the future (or at least not get worse), then it is prudent to prioritize research into the processes that will have high leverage over the future.

Further reading: Armstrong, *Smarter Than Us (https://intelligence.org/smarter-than-us/)*

Claim #4: Highly intelligent AI systems won't be beneficial by default.

We'd like to see the smarter-than-human AI systems of the future working together with humanity to build a better future; but that won't happen by default. In order to build AI systems that have a beneficial impact, we have to solve a number of technical challenges over and above building more powerful and general AI systems.

Alternative view: As humans have become smarter, we've also become more peaceful and tolerant. As AI becomes smarter, it will likewise be able to better figure out our values, and will better execute on them.

Short response: Sufficiently intelligent artificial reasoners would be able to *figure out* our intentions and preferences; but this does not imply (http://lesswrong.com/lw/igf/the_genie_knows_but_doesnt_care/) that they would execute plans that are in accordance with them.

A self-modifying AI system could inspect its code and decide whether to continue pursuing the goals it was given or whether it would rather change them. But how is the program deciding which modification to execute?

The AI system is a physical system, and somewhere inside it, it's constructing predictions about how the universe would look if it did various things. Some other part of the system is comparing those outcomes and then executing actions that lead towards outcomes that the current system ranks highly. If the agent is initially programmed to execute plans that lead towards a universe in which it predicts that cancer is cured, then it will only modify its goal if it predicts that this will lead to a cure for cancer.

Regardless of their intelligence level, and regardless of your intentions, computers do *exactly* what you programmed them to do. If you program an extremely intelligent machine to execute plans that it predicts lead to futures where cancer is cured, then it may be that the shortest path it can find to a cancer-free future entails kidnapping humans for experimentation (and resisting your attempts to alter it, as those would slow it down).

There isn't any spark of compassion that automatically imbues computers with respect for other sentients once they crosses a certain capability threshold. If you want compassion, you have to program it in.

Why this claim matters: A lot of the world's largest problems would be much easier to solve with superintelligent assistance — but attaining those benefits requires that we do more than just improve the capabilities of AI systems. You only get a system that does what you intended if you know how to program it to take your intentions into account, and execute plans that fulfill them.

Further reading: Bostrom, "The Superintelligent Will (http://www.nickbostrom.com/superintelligentwill.pdf)"

These four claims form the core of the argument that artificial intelligence is important: there is such a thing as general reasoning ability; if we build general reasoners, they could be far smarter than humans; if they are far smarter than humans, they could have an immense impact; and that impact will not be beneficial by default.

At present, billions of dollars and thousands of person-years are pouring into AI *capabilities* research, with comparatively little effort going into AI safety research. Artificial superintelligence may arise sometime in the next few decades, and will almost surely be created in one form or another over the next century or two, barring catastrophe. Superintelligent systems will either have an extremely positive impact on humanity, or an extremely negative one; it is up to us to decide which.

Did you like this post? You may enjoy our other Analysis (https://intelligence.org/category/analysis/) posts, including:

- Decision Theory (https://intelligence.org/2018/10/31/embedded-decisions/)
- Yudkowsky and Christiano discuss "Takeoff Speeds" (https://intelligence.org/2021/11/22/yudkowsky-and-christiano-discuss-takeoff-speeds/)
- Three Major Singularity Schools (https://intelligence.org/2007/09/30/three-major-singularity-schools/)
- Embedded World-Models (https://intelligence.org/2018/11/02/embedded-models/)
- ... and many more (https://intelligence.org/category/analysis/).

Pranav Budhwant

7 months ago

RE: Short response for Alternative view #2, for claim #2:
I feel this response misses the point -- the fact the building blocks of general intelligence was present in species far back in the evolutionary chain has no implication on how easy or difficult it is to program intelligence in computers. Couldn't it still be extremely complicated to replicate on computer systems?