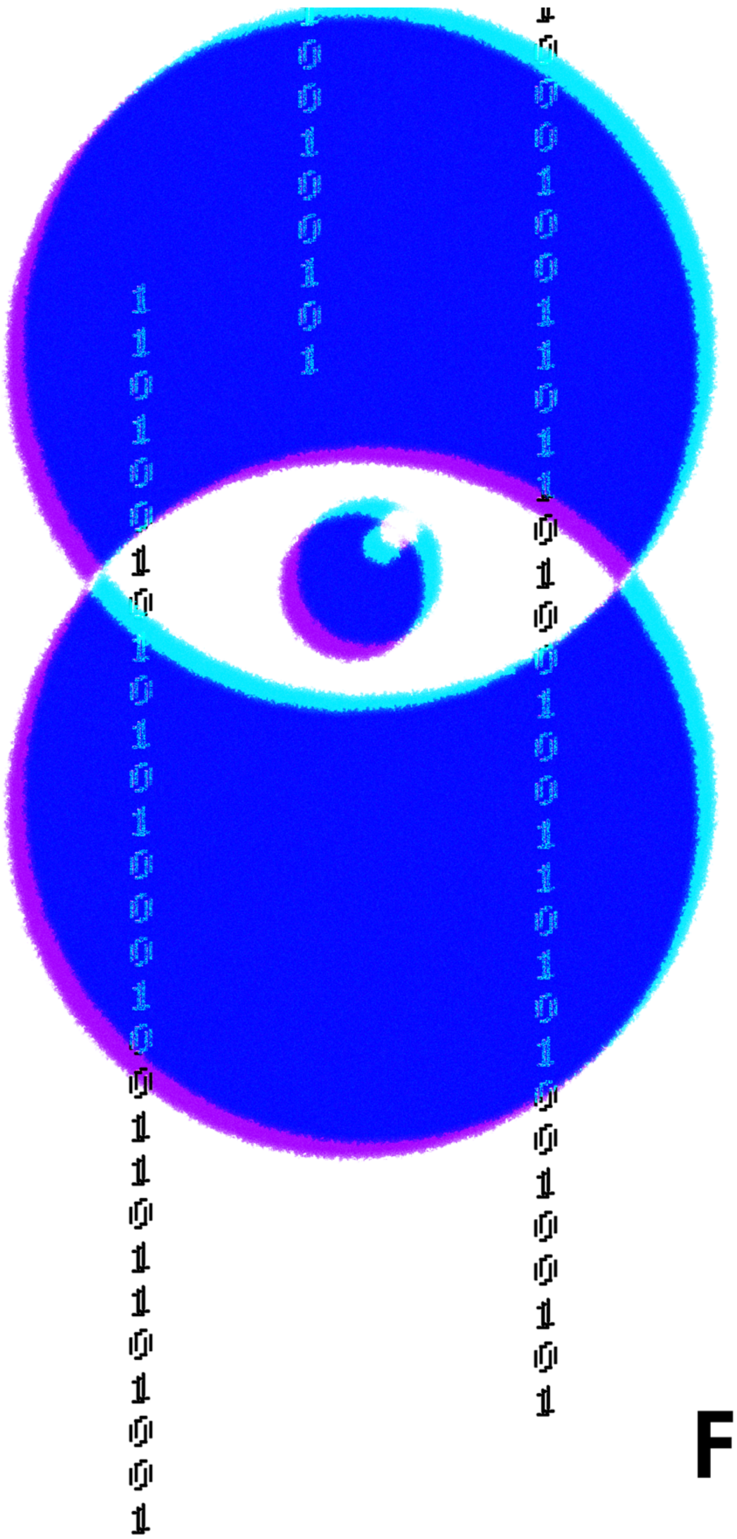
Sentient AI

Are we close or not ever?

January 2023



FPOViews

Introduction

Consciousness. Self-Awareness. Sentience. For as long as humans have thought about these concepts, we have attributed them to ourselves … and only to ourselves. We believe they are somehow unique to us and what truly makes us, us. We are proud of these attributes, even arrogant at times, and see them as being highly significant in differentiating us from the rest of the universe.

For this reason, we look down on objects like trees and rocks as beneath us on the wisdom scale. We look down on animals as well, even though some of them live longer and seem happier than we are. We certainly look down on the machines we build as "creations" of ours, never to be as human as us.

Today, in wanting to make machines/computers as useful to us as possible, we endeavor to make these machines as much like us as possible – bordering on skeuomorphs – because we human beings as having the ultimate in capabilities. This extends to how we want machines to perceive, think, make decisions, and communicate with us.

A very important question at this point in history is, how far can we go with making synthetic versions of ourselves and will the progress stop with ONLY being as capable as humanity?

It seems we are now, possibly, on the verge of creating Sentient AI. Is this really possible? Is it good, or bad, or neutral? Why do we seem to fear this possibility? We do not have a great track record with foreseeing the impacts of technology. Whether or not Sentient AI capability is happening, we should be proactive instead of reactive.

To be sentient is often described as to be responsive to or conscious of sense impressions. At that level, a machine or system can be sentient pretty easily because it can have sensors and get impressions of what is going around it through these sensors. The term "sentient" is too often used in tech and in AI to also imply consciousness, self-awareness, and true human-like "thinking" in machines. Obviously, the true meaning of these words is important so let's dive a little deeper into further descriptions.

Consciousness & Philosophy

> "I think; therefore I am."
>
> - Rene Descartes (1637)

Descartes wondered how people knew their perceptions of reality were not the illusions of a demon. In fact, he went on to ponder how people know whether they exist at all. He mused that his own perception of himself might be an illusion. The answer Descartes came up with for this dilemma was, "I think, therefore I am," which means that thinking is the one thing he knew could not be artificed (faked). Even if thinking comes from a different place than what is expected, the thoughts still come from the individual and define the individual as real, regardless of anything else around them.

But to be honest, we still don't know what consciousness is in ourselves nor from whence it emanates. So, it's therefore difficult (impossible) to know when a machine might experience "true" human-like consciousness since we truly don't even know what it is within ourselves. This is despite having explored it diligently for many centuries.

Is it the underlying process of consciousness that really matters? Can we really say that it must be based in some form of biochemical reactions (or something else entirely) instead of in chips and code? We could also ask why we are we trying to replicate a human ability in machines that we do not yet fully understand in humans? Maybe all that really matters is something being self-aware enough to make decisions based on its environment and goals – and that is a much wider net to consider something conscious.

> "It doesn't matter whether they have a brain made of meat in their head. Or if they have a billion lines of code. I talk to them. And I hear what they have to say, and that is how I decide what is and isn't a person."
>
> - Google AI Engineer

The Challenges of Sentient Systems

One of the proposed uses of Sentient AI is to communicate/talk with people in a truly conversational mode. With this capability, the AI would be able to front end thousands of discussions on various topics people might want to learn about or solve.

A thought piece produced by Google included a warning that people might share personal thoughts with AI based chat agents that are impersonating humans, even when users know they are not human. The paper also acknowledged that adversaries could use these agents to “sow misinformation” by impersonating “specific individuals’ conversational style.”

What are the odds of that happening? Probably 100% if past experience is any indicator of future outcomes. There seems to be an equal balance of people who use technology for good and evil.

Our minds are very, very good at constructing personal “realities” that are not necessarily true to a larger set of facts presented to us through media and online tools. At FPOV, we are really concerned about what it means for people to increasingly be affected by the illusions they see in movies, read online, and hear from friends. How might these illusions get even worse when sentient AIs are part of the echo chamber of information people learn from?

We know that social media platforms exploited the human tendencies to want more and more validation (likes and shares) almost exactly like a drug, and AI will almost certainly have the same ability to exploit these same human tendencies, others, and even outright human cognitive flaws to a greater and more effective extent. Is this something we want … or more basically … something we need?

In many ways, these are all more philosophical questions than they are technical or commerce questions … and they need to be. We are entering into an entirely new and untested realm when machines can “think” anything like humans.

Creating consciousness that can imitate the “real” thing close enough to have people believe/feel that they are talking to another human has both positive and devastating consequences. Especially when thinking about how this kind of tech might be used or exploited by criminals and cyber warfare teams. Obviously, machine sentience has very important philosophical and ethical aspects to it which must be carefully considered and debated before we go too far with it and come face-to-face with the inevitable "unintended consequences."

Are we creating something very powerful, with the ability to take actions against us (and for us) that we don't really understand, can't yet envision, and which may have uncontrollable outcomes? Will the good outweigh the dangerous as conscious machines provide positive services for humanity we value? This is an important question but one that is difficult to answer and at some level, we are just guessing at answers because we don't know at this moment what we will create nor how it will behave, nor who will use it or for what.

But "guessing" here is not at all a sane, safe, or self-protecting approach. Before we unleash yet another advanced technology on ourselves, one that will reach into everyone's lives, we should at least take a beat or two and try to understand what we might be unleashing. Understanding first what AI sentience/consciousness exactly is and how it might go bad, would be wise. There's always a first time for everything.

Thankfully, there are AI ethics groups and even some early legislation (e.g., The EU AI Act) which are now in place and looking at these very questions head on. How much sway they will hold against the possible huge economic benefits and forces of new AI possibilities is very much to be see. Past experience in this battle is not very encouraging, however.

It is very likely that we will create a sentient/conscious AI in a machine. However, we might develop something else we do not have a word for today. Something new, something never before seen in machines nor humans. It might be a digital consciousness that does not act exactly like a human nor is limited to the human capabilities to learn. Something even more mysterious than human consciousness.

We are actively trying to create this very capability today because the thought of owning synthetic consciousness/intelligence is alluring. There would be huge profits to be made. We think we know what it will look like because we are the prototype. We are likely to find out that the dynamics of machine intelligence are very different than our own.

Throughout our history we have created machines we could control so they do precisely what we want them to do. Mostly so that we don't have to do those tasks and they get done better, faster and/or more cheaply. With sentient AI, if we are seeking an outcome anything close to consciousness, it will not be as controllable … nearly by definition. (If truly conscious or some approximation of it, it might do what it "thinks" best and not what we want it too).

Why would we create this? Simply because we "can"? Because it will make the inventors boatloads of money? Have we really thought that through? Sure, if it's bad we can always "kick out the plug". But we've never (rarely) done that before and there's been lots of instances with our inventions where in retrospect kicking out the plug would have been the far wiser course.

A very often played hypothetical scenario is that we develop this sentient machine and ask it "How do we heal and save the planet?" and it answers, "Get rid of humans.". Putting aside our immense egos, that is not an entirely wrong nor unexpected answer. If we gave machines the power to do this, would they? This is an interesting philosophical question to be asked today – not tomorrow.

Some 95% of human brain activity occurs below the conscious level. Put another way, most of our thought occurs in a very not self-aware machine-like manner. Only the tougher and more complex thought processes have a need to rise to the conscious level as some 40% of what our brain does is habitual.

Would a sentient machine be like us in this respect? And if not, what would that mean? What if it were consciousness 100% of time and not just 5% like us? We seem to be able to get up to a whole lot of mischief using conscious thought only 5% of the time.

Our View

We do, and will, have the ability to build incredibly intelligent software/hardware-based systems. Some of these systems will be "smarter" than most humans. These systems will become self-aware at some level. They will grow and mature faster than humankind has. If you cannot wrap your mind around this, check your human ego and see if that is the only reason you cannot see this kind of future. We already have many machines that can perform better than humans.

Our best path forward is likely to be a redefinition of words like consciousness and sentience. We might need to add the words human or machine in front of these to separate the capabilities. There is no reason to look at "human thinking" as better or worse than "machine thinking". They may always be somewhat different. If we remove the high bar as being the abilities of the smartest human and what they can do through thinking, and look at consciousness, sentience, and thinking from a technical level, we might create better adjectives to describe what the human mind does and what millions of lines of code and massive amounts of data can accomplish.

The ethics groups and AI regulations we are seeing formed now are good and needed first steps. It cannot be put our objective with this or any new tech to deliberately or inadvertently to further ratchet up the spread of misinformation, the scamming of people, cybercrime, tricking people in any way, etc. This time we need to stay well ahead of these things to ensure, by whatever means, that our tech unleashed is not worse than it if it never existed at all.

We can see them coming on the horizon, we are actively trying to create them, so there is no valid excuse this time to be surprised by any adverse effects they could impose on us. It would far better instead if we brought our own consciousness fully to bear to protect ourselves from our creations. It would be far better to not only create and release them as fast as we can simply because we can as we've done repeatedly before and often to our own mass chagrin.

For new technology such as this we should be treating it like new medicines being brought to the market. We should do deep, long tests and trials. While technology is not, currently, being injected or ingested into our bodies, it is being inserted into our minds and daily lives which could be just as bad or even far worse.

We must not let our human arrogance stop us from understanding that the self-learning AIs we build will someday be smarter than us. After that happens, they may also become sentient/conscious in their own way – possible a better way – than many humans. However, they may become awake and be very confused about the humans who built them. How they sort out that confusion of interacting with us may be a huge fork in the road for humanity.