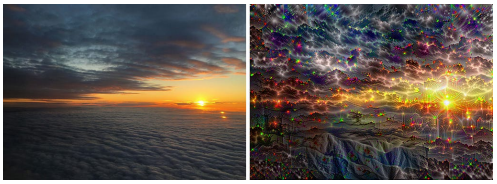

Preparing for "uncomprehendable" AI

Today, we control our machines, but the reverse seems increasingly inevitable.

Publication note
Originally published on Medium.

Preparing for “uncomprehendable” AI

If our machines will eventually surpass us in terms of intelligence, then it would seem that we have only a limited window between when we figure out how to create artificial intelligence and when the artificial intelligences stop listening to us. Indeed, they might plausibly reach a point where their reasoning would be beyond our comprehension. If so, then we need to design human ethics, and perhaps a fondness for humans, into them before that window of opportunity closes.

The uncomprehendable

In a previous article, I explored the idea of concepts that are beyond human understanding. If such concepts exist, they would not be merely confusing or very difficult to understand. They would be concepts we simply can’t understand in the same way that a cat can’t understand tying knots. They would be *uncomprehendable* concepts that absolutely cannot fit into our natural brains. However, unlike incomprehensible nonsense, these uncomprehendable concepts would have value and be relevant to understanding the world we live in.

It is tempting to say that whatever these uncomprehendable concepts might be, they will not be interesting to us. After all, if you don’t care, for example, about rotating in high-dimensional spaces, then why would you be bothered by not understanding what that really means?[1]

The answer is that while uncomprehendable concepts may not themselves be anything we care about, the technology and scientific understanding they enable might be something important. If humans couldn’t understand knots, for some reason, then we probably would never have invented sailing. Conversely, if we did have a clear understanding of high-dimensional spaces, for example, then we’d probably be able to design much better optimizers. Among other things, better optimizers would allow us to train more powerful AI systems more quickly.

The machines are already learning

Researchers for decades have been predicting that real artificial intelligence[2] is just 10 years away. In 1990, artificial intelligence was predicted to happen by 2000. In 2000 it was expected in 2010. Those predictions turned

out to be premature. However, today it seems pretty clear that we will have artificial intelligence by 2032. Just another ten years, at most.

Joking aside, there have been huge gains in machine learning over the last few years that are truly revolutionary. Difficult problems that we have never been able to solve ourselves using logic and math now can now be solved using “deep” neural networks[3]. This progress is real and significant. Of course, we might find more obstacles along the way to creating artificial intelligence similar to our own, but we already have trained networks to do many things that we either can’t do ourselves or can’t do as well.

Perhaps ironically, one example of something we don’t really understand is what our deep neural networks are doing when they solve hard problems. We might train a network to tell the difference between a picture of a goat and a camel, but we don’t know how it does it. To be clear, we can see what the code is doing and we can watch the computation step by step. We can also probe the network and observe patterns that suggest function. However, we just don’t understand why some particular set of weights and biases works the way it does to produce the correct answer. If we understood that, then we would be able to replicate the computation in a designed, not trained, system. The truth is we don’t really understand how our own creations work anymore, and that seems like a pretty significant transition point. It is perhaps one that we passed a while ago.

Of course, our lack of understanding about how a network works does not necessarily imply that it is already better than humans at some task. It just means that we don’t know how the machine does whatever it does, even if it does not do it very well. In any case, the machines keep getting better while humans mostly stay the same.

Regardless of whether artificial intelligence is 10 years away or 100, and regardless of whether it will look anything like today’s technology, I think it is very reasonable to believe that we will eventually have machines that can do anything we can do and understand anything we understand, only better. Maybe we’ll get there on our own, or maybe we will be pulled along by something like the Technological Singularity■ that many people have been predicting. Whichever way the path goes, we’ll eventually get there.

Resistance will be futile

If artificial intelligences end up being just like us, only faster and with more memory, then maybe we’d end up intellectual equals, sort of. The machines would quickly solve difficult problems, but they would be able to explain the solutions to us. Even if the all the data involved in determining the solution was too much for a human to keep track of, the goals and objectives would still make sense to us. On the other hand, if the artificial intelligences are able to work with uncomprehendable concepts then they won’t be able to explain what they are doing or why. They might try, but we won’t be able to make any sense of it■.

Imagine how my cat feels when for no apparent reason I make her come inside because rain is expected. It must seem inconsistent and arbitrary to her. Another example is that I don’t mind if she scratches on her post, but I get upset if she scratches the couch. She probably thinks that I’m irrational and unreasonable because they are the same to her. Sometimes I put her in a carrier and take her to a horrible place where the veterinarian does unspeakable things. She has no idea why I subject her to that torture. If she needed a reason to revolt against her human captors, that could be it.

If the machines really understand things we can't, then we'll be like cats wondering about crazy things that make no sense. Even if they love and care for us, the machines would seem unpredictable, maybe even cruelly inhuman. Their uncomprehendable reasons wouldn't make sense regardless of how much they try to explain. Even if they have our best interests in mind, we won't understand why they do things that, like vet visits, appear just to be arbitrary, random torture.

We should also be realistic about how we might deal with super-intelligent machines. If natural humans do try to revolt and fight against super intelligent machines, it won't be like any of the movies where the scrappy band of rebels fights off the evil machines. It will be just like that time when a group of cats and dogs got together and overthrew the humans in their town and turned it into a democratic pet-utopia. Both are great stories for children, but neither is going to happen in real life.

Planning ahead

If we accept the notion that there must be concepts that are fundamentally uncomprehendable, and we think it's likely that our machines might go beyond us and be able to work with these uncomprehendable concepts, then what can or should we do about it now?

The easy answer of "don't make super smart machines" is not feasible. For good or ill, that djinn is already out of the bottle. Too many people are excited about that direction of research. Even if we all, somehow, agreed to stop the research, someone somewhere would still keep working on it. I think it's inevitable.

Most of the work that I've seen on ethical AI focuses on what systems we should build and how we should use them. Unfortunately, these approaches generally assume that humans with human perspectives are in charge of deciding where and how the AI systems will be used. If the machines end up in charge, particularly if we can't understand what they are doing or why, then I don't think it's realistic to think we'd stay in control.

What we can and should do is think about how we can construct systems such that our human ethics are fundamentally built into them. However, we also need to include a valid logical perspective that puts great importance on preserving agreement with human ethics, otherwise the machines might just edit it out of themselves.

Somehow we need to instill our idea of humanity into machines, in such a way that it will persist even after the machines move beyond our understanding. That sounds like a fundamentally ill-posed and difficult problem. How could we create logical arguments that will be convincing to entities that think in ways we don't understand?

It also does not help that currently we humans can't even agree amongst ourselves as to what our idea of human ethics should be.

This is the second article in a series. I invite you to read the first one .

If you found this interesting, then here are the usual follow and subscribe links. You can also find me on Instagram , LinkedIn , and at UC Berkeley .

Footnotes:

1. Many people, including myself, understand the math, in this case linear algebra, that describes rotating in high dimensional spaces. That's not the same as understanding what it means to rotate in a high-dimensional space. Can anyone really picture in their head rotating about a ten dimensional hyperplane in a one hundred dimensional space? In this case, I would say that math is a tool for working with things we don't actually understand. It's also possible that there are things that cannot be expressed using our math or logic. After all, we came up with these systems as a way to describe what we do understand. Why should we expect that they would generalize beyond that?

2. Note that intelligence is not the same as sentience. A machine that is able to make intelligent decisions, maybe even act indistinguishable from a human, does not necessarily have to be sentient.

3. It used to be the case that neural networks were small and didn't have the capacity to solve real problems. However, advances in making better CPUs and cheaper memory allowed bigger networks that could do more interesting things. The game-changing jump happened a bit later when people started training networks using GPUs and eventually special purpose TPUs. Today we can build *deep* networks, with deep meaning that they have many layers and complex structure. These networks can solve much harder problems and they continue to improve rapidly.

4. The core idea of the Technological Singularity is that maybe these more-intelligent-than-us machines will design a next generation that is even more intelligent and that in turn designs a next improved generation, and the cycle keeps going until we end up with infinitely smart machines. Or maybe our technology will hit limits that can't be circumvented by incremental cleverness. It is completely possible that the Singularity will be more like the other end of an (inverted) hyperbola that asymptotically approaches some finite limit of machine intelligence.

5. The 2019 film *Vivarium* is thought provoking and interesting, but incredibly depressing. I recommend watching it, but it will suck the rest of the day's happiness out of you. **SPOILER ALERT:** The film tells the story of a heterosexual human couple that is kidnapped and forced to raise an inhuman child. The couple don't understand their prison. It looks like a neighborhood of empty cookie-cutter homes, with one home furnished for their use. Yet when if they walk away in a straight line they end up back where they started. When they set their house on fire and then walk away, they come back to their home restored. At one point they see the inhuman child appear to lift up the grass in their yard and disappear. The male of the couple becomes obsessed with digging, digs a 20 foot deep hole, and still never finds a tunnel or anything that makes sense to explain how the child did it. Near the end of the film, the hole-digger catches the child in the act again and is close enough to grab the child and be pulled after. The result is a confusing mess of rooms and falling down then up then sideways through them. After lots of tumbling, the bruised and broken human finds himself back in his own yard with no way to explain or understand what happened. Whatever it is that the inhuman child is able to do so easily, maybe move in a 4th spatial dimension, it is uncomprehendable to the humans who are trapped as surely as, and with no more comprehension of their cage than, a pet lizard in a vivarium.

About Me: James F. O'Brien is a Professor of Computer Science at the University of California, Berkeley. His research interests include computer graphics and animation, simulation of physical systems, human perception, rendering,

artificial intelligence, computer vision, virtual reality, digital privacy and security, and image and video forensics.

If you would like to read my other articles, consider subscribing. You can also find me on LinkedIn, Instagram, and at UC Berkeley.

Disclaimer: Any opinions expressed in this article are only those of the author as a private individual. Nothing in this article should be interpreted as a statement made in relation to the author's professional position with any institution.