Controlling what you don't understand

How do we control AI systems after they have evolved beyond human understanding?

Publication note
Originally published on Medium.

Controlling what you don’t understand

I’ve previously written about how there must be concepts beyond possible human comprehension. Concepts that are not just difficult to understand or requiring extensive study. Rather, concepts that are completely inaccessible to the human mind. While we typically label things that can’t be understood as “incomprehensible,” implying nonsense or illogicality, I propose a different term for concepts that have consistent logic and relate in some meaningful way to our universe, yet they cannot be understood by a natural human: “uncomprehendible.” These are ideas that do make sense but evade human understanding in the same way that comprehending calculus is out of a cat’s reach.

Envisioning what type of thing an uncomprehendible idea might be is inherently ill-posed. When contemplating the existence of the uncomprehendible, one attempts to imagine some clear example that would be uncomprehendible and invariably fails. It is nearly a tautology that one cannot clearly envision what one cannot comprehend. We can point to things that we don’t understand, for example structures in the weights of deep neural networks, but we still don’t actually identify the specific uncomprehendible concept that is not understood. So for deep neural networks, we imagine the existence of unimaginable, complex structures abstractly with no answer to what specifically they are and how they are uncomprehendible.

Confronted with the concept of the uncomprehendible, a scientist might argue that logical and mathematical tools allow us to manipulate and reason about these uncomprehendible concepts, even if we can’t intuitively grasp them. However, even if we accept the questionable assumption that the intellectual tools we’ve built are sufficient, this perspective oversimplifies the practical use of such tools.

Consider the simple straight line, the shortest distance between two points. Most intuitively grasp this fundamental concept and its implications. Even young children can see that a curved path between two points is longer than the straight one and that making the curve straighter shortens it. However, proving mathematically that the shortest path between two points has zero curvature requires advanced mathematical tools, like the calculus of variations, that are typically learned only after at least a decade of mathematical study beyond when the child intuitively solved the same problem with crayons.

Math and logic are powerful tools that we can use to search for truths, but intuition tells us where in the vast space of possibilities we should look.

Consider the impact of intuitive understanding on technological advancements. An intuitive understanding of linear geometry based on straight lines allows a skilled craftsperson to visualize the construction of a bookcase or building. Imagine if humans lacked this intuitive grasp and instead needed advanced mathematical tools just to design a simple bookcase. So many technologies we regard as simple and basic would instead be hideously complex.

Conversely, imagine a hypothetical species just like us, but also with a native and intuitive understanding of some complex mathematical concepts that we can only manipulate with formal tools. Our PhD dissertations on those topics would be their child-play and those children's school assignments would to us be uncomprehendible.

Human cognition, constrained by evolutionary biology, is constant. In contrast, our machines continue to advance at an accelerating rate. Regardless of whether these machines do or don't ever become conscious, at some point when we put a machine to a task we will no longer be able to understand what it is doing or why.

As we continue inevitably toward a future where our technology surpasses our understanding, we need to somehow design systems that will remain aligned with our best interests as they evolve. We must develop strategies for ensuring these systems continue to serve us effectively, even as their inner workings become increasingly opaque and their actions increasingly inscrutable.

The assumption that advanced, logical reasoning would lead to decisions that we humans are happy with overlooks the essential humanity of our decision-making processes. Humans often prioritize values and principles over pure logic, leading to decisions that may appear irrational from a a purely analytical perspective, yet are nevertheless deeply significant to us. Paradoxically, it is these deeply significant things where we have the most trouble explaining clear reasoning to support our opinions and where we have the most trouble agreeing with each other.

Given these challenges, can we reasonably expect to implement in our machines humane decision making when we can't even clearly define for ourselves what that means? Today we can review what our machines do and see if their output conforms to our human sensibilities, but at some point we won't be able to review what we can't understand. What can we do about it?

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.