What Does It Mean to Align AI With Human Values?

Making sure our machines understand the intent behind our instructions is an important problem that requires understanding intelligence itself.

By Melanie Mitchell

Jun 11, 2024 06:44 PM · 7 min. read · View original

Many years ago, I learned to program on an old Symbolics Lisp Machine. The operating system had a built-in command spelled “DWIM,” short for “Do What I Mean.” If I typed a command and got an error, I could type “DWIM,” and the machine would try to figure out what I meant to do. A surprising fraction of the time, it actually worked.

The DWIM command was a microcosm of the more modern problem of “AI alignment”: We humans are prone to giving machines ambiguous or mistaken instructions, and we want them to do what we mean, not necessarily what we say.

Computers frequently misconstrue what we want them to do, with unexpected and often amusing results. One machine learning researcher, for example, while investigating an image classification program’s suspiciously good results, discovered (opens a new tab) that it was basing classifications not on the image itself, but on how long it took to access the image file — the images from different classes were stored in databases with slightly different access times. Another enterprising programmer (opens a new tab) wanted his Roomba vacuum cleaner to stop bumping into furniture, so he connected the Roomba to a neural network that rewarded speed but punished the Roomba when the front bumper collided with something. The machine accommodated these objectives by always driving backward.

But the community of AI alignment researchers sees a darker side to these anecdotes. In fact, they believe that the machines' inability to discern what we really want them to do is an existential risk. To solve this problem, they believe, we must find ways to align AI systems with human preferences, goals and values.

This view gained prominence with the 2014 bestselling book *Superintelligence* by the philosopher Nick Bostrom, which argued in part that the rising intelligence of computers could pose a direct threat to the future of humanity. Bostrom never precisely defined intelligence, but, like most others in the AI alignment community, he adopted a definition later articulated (opens a new tab) by the AI researcher Stuart Russell as: "An entity is considered to be intelligent, roughly speaking, if it chooses actions that are expected to achieve its objectives, given what it has perceived."

Bostrom based his view of AI's risks on two theses. The first is the orthogonality thesis, which states, in Bostrom's words, "Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.” The second is the instrumental convergence thesis, which implies that an intelligent agent will act in ways that promote its own survival, self-improvement and acquisition of resources, so long as these make the agent more likely to achieve its final goal. Then he made one final assumption: Researchers would soon create an AI superintelligence — one that “greatly exceeds the cognitive performance of humans in virtually all domains of interest.”

For Bostrom and others in the AI alignment community, this prospect spells doom for humanity unless we succeed in aligning superintelligent AIs with our desires and values. Bostrom illustrates this danger with a now-famous thought experiment: Imagine giving a superintelligent AI the goal of maximizing the production of paper clips. According to Bostrom’s theses, in the quest to achieve this objective, the AI system will use its superhuman brilliance and creativity to increase its own power and control, ultimately acquiring all the world's resources to manufacture more paper clips. Humanity will die out, but paper clip production will indeed be maximized.

If you believe that intelligence is defined by the ability to achieve goals, that any goal could be "inserted" by humans into a superintelligent AI agent, and that such an agent would use its superintelligence to do anything to achieve that goal, then you will arrive at the same conclusion (opens a new tab) that Russell did: "All that is needed to assure catastrophe is a highly competent machine combined with humans who have an imperfect ability to specify human preferences completely and correctly."

It's a familiar trope in science fiction — humanity being threatened by out-of-control machines who have misinterpreted human desires. Now a not-insubstantial segment of the AI research community is deeply concerned about this kind of scenario playing out in real life. Dozens of institutes have already spent hundreds of millions of dollars on the problem, and research efforts on alignment are underway at universities around the world and at big AI companies such as Google, Meta and OpenAI.

What about the more immediate risks posed by non-superintelligent AI, such as job loss, bias, privacy violations and misinformation spread? It turns out that there's little overlap between the communities concerned primarily with such short-term risks and those who worry more about longer-term alignment risks. In fact, there's something of an AI culture war, with one side more worried about these current risks than what they see as unrealistic techno-futurism, and the other side considering current problems less urgent than the potential catastrophic risks posed by superintelligent AI.

To many outside these specific communities, AI alignment looks something like a religion — one with revered leaders, unquestioned doctrine and devoted disciples fighting a potentially all-powerful enemy (unaligned superintelligent AI). Indeed, the computer scientist and blogger Scott Aaronson recently noted (opens a new tab) that there are now “Orthodox” and “Reform” branches of the AI alignment faith. The former, he writes, worries almost entirely about “misaligned AI that deceives humans while it works to destroy them.” In contrast, he writes, “we Reform AI-riskers entertain that possibility, but we worry at least as much about powerful AIs that are weaponized by bad humans, which we expect to pose existential risks much earlier.”

Many researchers are actively engaged in alignment-based projects, ranging from attempts at imparting principles (opens a new tab) of moral philosophy to machines, to training large language models (opens a new tab) on crowdsourced ethical judgments. None of these efforts has been particularly useful in getting machines to reason about real-world situations. Many writers have noted the many obstacles preventing machines from learning human preferences and values: People are often irrational and behave in ways that contradict their values, and values can change over individual lifetimes and generations. After all, it's not clear whose values we should have machines try to learn.

Many in the alignment community think the most promising path forward is a machine learning technique known as inverse reinforcement learning (opens a new tab) (IRL). With IRL, the machine is not given an objective to maximize; such "inserted" goals, alignment proponents believe, can inadvertently lead to paper clip maximizer scenarios. Instead, the machine's task is to observe the behavior of humans and infer their preferences, goals and values. In recent years, researchers have used IRL to train machines to play video games (opens a new tab) by observing humans and to teach robots how to do backflips (opens a new tab) by giving them incremental feedback from humans (people viewed short clips of a robot's various attempts and chose the one that looked best).

It's unclear whether similar methods can teach machines the more subtle and abstract ideas of human values. The writer Brian Christian, author of a popular science book about AI alignment (opens a new tab), is optimistic: “It’s not such a stretch to imagine replacing the nebulous concept of ‘backflip’ with an even more nebulous and ineffable concept, like ‘helpfulness.’ Or ‘kindness.’ Or ‘good’ behavior.”

However, I think this underestimates the challenge. Ethical notions such as kindness and good behavior are much more complex and context-dependent than anything IRL has mastered so far. Consider the notion of “truthfulness” — a value we surely want in our AI systems. Indeed, a major problem with today’s large language models is their inability to distinguish truth from falsehood. At the same time, we may sometimes want our AI assistants, just like humans, to temper their truthfulness: to protect privacy, to avoid insulting others, or to keep someone safe, among innumerable other hard-to-articulate situations.

Other ethical concepts are just as complex. It should be clear that an essential first step toward teaching machines ethical concepts is to enable machines to grasp humanlike concepts in the first place, which I have argued is still AI’s most important open problem (opens a new tab).

Moreover, I see an even more fundamental problem with the science underlying notions of AI alignment. Most discussions imagine a superintelligent AI as a machine that, while surpassing humans in all cognitive tasks, still lacks humanlike common sense and remains oddly mechanical in nature. And importantly, in keeping with Bostrom’s orthogonality thesis, the machine has achieved superintelligence without having any of its own goals or values, instead waiting for goals to be inserted by humans.

Yet could intelligence work this way? Nothing in the current science of psychology or neuroscience supports this possibility. In humans, at least, intelligence is deeply interconnected with our goals and values, as well as our sense of self and our particular social and cultural environment. The intuition that a kind of pure intelligence could be separated from these other factors has led to many failed predictions (opens a new tab) in the history of AI. From what we know, it seems much more likely that a generally intelligent AI system’s goals could not be easily inserted, but would have to develop, like ours, as a result of its own social and cultural upbringing.

In his book *Human Compatible*, Russell argues for the urgency of research on the alignment problem: “The right time to worry about a potentially serious problem for humanity depends not just on when the problem will occur but also on how long it will take to prepare and implement a solution.” But without a better understanding of what intelligence is and how separable it is from other aspects of our lives, we cannot even define the problem, much less find a solution. Properly defining and solving the alignment problem won’t be easy; it will require us to develop a broad, scientifically based theory of intelligence.