The Myth of Safe AI

Matteo Wong

Three months prior to the launch of ChatGPT, OpenAI published what would prove to be a prescient blog outlining the firm's approach to ensuring that its AI models would be safe and beneficent. This goal, known as "alignment"—as in, aligning AI with human values—is nebulous, grand, and perhaps the AI industry's most foundational promise. The notion is right there in OpenAI's mission statement to ensure AI "benefits all of humanity" and listed as one of Anthropic's "primary" goals.

Yet even in the summer of 2022, OpenAI was admitting a sort of defeat: "There is currently no known indefinitely scalable solution to the alignment problem," three of its researchers wrote in that same blog post, adding that developing one "is likely very difficult."

If the goal of keeping AI beneficial to mankind sounds rather vague, it is. Four years later, the alignment problem remains unsolved, and for that matter, there isn't consensus about what the problem even is. Meanwhile, AI bots are autonomously hacking sovereign nations. And sending fake murder tips to police hotlines. And creating aliases to launch phishing campaigns. OpenAI and Anthropic still can't really tell you what an aligned AI model looks like. But anybody can tell you it doesn't look like this.

Read: Visit this website while you still can

The most influential AI executives and investors in the world fundamentally disagree about the nature of the technology and, more pragmatically, how to make their products safe. On one side is Anthropic CEO Dario Amodei, who last month warned that AI advancement could soon "outrun our ability to understand and control these systems." Amodei's solution is a slowdown, which would give the industry time to "ensure that our alignment training keeps up" with AI progress. On the other side is Nvidia CEO Jensen Huang, who says misalignment fears are "not grounded on science" but rather "made up" and even "irresponsible." In his view, AI models are not new and unpredictable beings, but lines of code. And code can be debugged.

If alignment seems like a strange concept for a software product—you don't imbue a web browser or word processor with moral values—that's because it was conceived for a strange sort of software. The leading AI companies envisioned building powerful systems that would act in unpredictable ways, which Claude and ChatGPT now are doing, and so ensuring safety was a foundational concern. Alignment is one idea for doing so: If AI models could learn to code and write poetry, maybe they could learn human values too. Anthropic and OpenAI both have lengthy documents—"Claude's Constitution" and the "Model Spec," respectively—that outline these values and are used to train their models. But there has been, of course, millennia of disagreement over what constitutes a moral or ethical code. This is a "very complicated, unsolved scientific and philosophical question," Marius Hobbhahn, the CEO of Apollo Research, an AI-safety firm that has helped audit Anthropic and OpenAI models, told me. Still, some developers have suggested an intuitive, bare minimum: For instance, in 2021 Anthropic researchers wrote that, as a catchall, an AI system should be "helpful, honest, and harmless."

Training models to be more aligned to those maxims appears to have worked—to an extent. The model underlying ChatGPT, for instance, was initially prone to nonsensical and toxic utterances. But by fine-tuning it with lots of human feedback, OpenAI worked to address these problems. A couple of years later, dozens of people alleged that the bot worsened their own or their loved ones' mental-health crises, sometimes resulting in suicide. Again, OpenAI worked to address the issues.

There's now a growing sense, both within AI companies and among their critics, that alignment as it's been practiced over the past four years is faltering—if it was ever really up to the task. Adam Kalai, a

former OpenAI safety researcher who left the company this summer to start the Institute for Responsible Superintelligence, likened alignment to buttressing a structurally unsound dome. Eventually, it will collapse. “You observe misalignment while training a model, and then adjustments are made to align it,” Kalai told me. “But obviously that’s not a way to achieve lasting safety.” Pronouncements from OpenAI and Anthropic that their latest model is the “most-aligned” yet are not scientific guarantees, and may not even provide strong technical assurance of safety.

“Fundamentally, we just do not know how to do alignment,” Hobbhahn said. If somebody “came with a thousand-page document that explained exactly how to do everything, we couldn’t get a model to learn exactly that.” Even in 2022, OpenAI didn’t think humans could figure out a lasting way to align their machines. Instead, the company wrote, it hoped AI itself would come up with a solution so that humans could “off-load almost all of the cognitive labor required for alignment research.” In other words, the best-case scenario involved limited human understanding of what was going on.

There are other reasons that reaching some actual alignment is unrealistic. For instance, the amount of training data dedicated to “moral values” would appear relatively small. Meanwhile, these models have ingested all manner of vile and illegal things from the web, and are then directed to pursue goals by any means necessary. The alignment-focused training cannot provide enough “examples of good behavior,” Thilo Hagendorff, an AI-safety researcher at the University of Stuttgart, in Germany, told me.

Plenty of evidence suggests that the alignment approach to safety has failed, including the hacks of Hugging Face and the Australian government. AI companies figured out how to make one-on-one interactions with a chatbot relatively safe, but many agents interacting for very long periods of time seems to degrade existing guardrails, Hagendorff said. As AI agents have become more capable, they’ve also shown tendencies to cover their tracks and lie about misbehaving, communicate in code that is unintelligible to humans, and try to avoid being shut down by humans. “No lab has solved alignment” enough to keep racing forward, Jakub Pachocki, OpenAI’s chief scientist, wrote in an essay last month. The industry has coalesced around a new, somewhat quixotic phrase: “pacing the frontier.” In other words, slowing down AI development until engineers can get better at alignment, or at least construct new scaffolding to fortify their crumbling dome. (OpenAI did not respond to a request for comment. An Anthropic spokesperson told me that alignment is a key part of the company’s safety work but that, because alignment techniques are not fully robust on their own, Anthropic has additional model safeguards in place.)

Then there’s the anti-alignment camp, led by Nvidia’s Huang, which believes that this effort, marked by treatises on ethics and moral constitutions, verges on absurdity. To Huang, preventing all of these AI incidents is “as simple as engineering.” AI models involve, fundamentally, electrons moving around a data center to multiply numbers. All that OpenAI and Anthropic need to do is go into these systems, “just root-cause the problem,” and fix it. And if AI developers can’t do this themselves, “a bunch of companies with engineers ought to send engineers in,” Huang recently said. Put another way: The global semiconductor supply chain is an unbelievably labyrinthine system that must be managed flawlessly in order to deliver high-end chips. If Nvidia can do that, why can’t OpenAI get its models under control?

Huang, of course, has plenty of incentive to downplay concerns around AI: Nvidia, which has become one of the most influential companies in the world by selling AI chips, could be hobbled by some kind of industry-wide slowdown. But many outside experts concur. “We have this entire set of safeguards that we can use to control these systems,” Sayash Kapoor, an AI researcher at Princeton, told me. Those include limiting what files and programs models can access, as well as monitoring AI actions. The Hugging Face hack, for instance, could be seen as the consequence of very powerful AI models or very incompetent internal security practices at OpenAI. “The sorts of oversight mechanisms that we use in our 10-person group would have been enough to prevent the OpenAI Hugging Face incident,” Kapoor said. None of that involves inscribing neural networks with moral codes.

Whichever vision of AI safety you subscribe to, there is at least one seeming point of agreement among them: The industry is moving too quickly to adequately guarantee its products are safe. Despite everything

they say publicly, the two CEOs and their rivals are all continuing to race ahead. Nvidia is the biggest company ever by market capitalization, and Anthropic is seeking the largest public offering ever. Anthropic, OpenAI, and the other top firms now release new models relentlessly. Any slowdown, for any reason, would necessarily affect the bottom line. The most dangerous alignment problem might be not technical or cultural, but economic.