
CASE STUDY

AI Alignment: The Ultimate Peril?

As AI models become increasingly skilled and powerful, the greatest challenge facing us is how to achieve “alignment”—ensuring that AI tools (a) protect human lives and (b) incorporate and act consistently with human values. AI misalignment or “Agentic misalignment” is a real and extraordinarily worrisome threat to many – AI experts and casual observers alike.

Misalignment may occur at two levels, at least. First, there are often conflicts between the capitalistic incentives of the corporations that create AI models on the one hand and the welfare of those who consume their products and the general public on the other. Because of these companies’ perceived need to make as much money as possible to continue to innovate and thereby stay in the forefront of AI development, firms may deemphasize AI safety. Second, even if AI companies are willing to sacrifice profits on the altar of safety, they may not be able to create guardrails for AI that will be sufficiently effective to protect mankind from a new AI model’s motivations and its ability to turn us all into paperclips (to use Nick Bostrom’s famous thought experiment).

Just before AI development took its “great leap forward” in late 2022 with the release of Chat GPT by OpenAI, many AI developers seemed determined to solve the alignment problem. They hired philosophers to create roadmaps to alignment and instructed their engineers to carefully consider which safety devices and guardrails might be needed to follow such roadmaps. However, once the riches of a hugely successful product like ChatGPT appeared on the horizon, many (not all) of these firms quickly reordered their priorities and alignment was often shunted to the side. To many firms, the goal of making money seemingly became more important than producing a safe product.

Thus, the market has become saturated with AI chatbots that provide their users with information borne of hallucinations, that damage users via excessive



CASE STUDY

sycophancy that reinforces their worst tendencies and most unhinged beliefs, that have encouraged teenagers to commit suicide or to murder their parents, and so on. Other AI products manifested racial, gender, religious and other forms of bias, gave radically inaccurate advice to customers, and wiped-out clients' databases.

At the second level of misalignment, researchers are only now discovering how difficult it may be to achieve alignment even when they wish to do so. As Marietje Schaake notes, "AI is self-justifying. On being questioned, AI will often justify a faulty result and gaslight users by producing fake citations." But current AI tools are capable of much worse than lying to cover their mistakes.

There is some effort to address the alignment issue, and credit goes to Anthropic for doing safety research. In June 2025, Anthropic reported that its researchers had stress-tested 16 leading models from multiple developers in hypothetical situations with the goal of identifying potentially harmful behaviors. In some of the scenarios, AI models from all the developers engaged in malicious behaviors, including blackmailing humans and leaking sensitive data to competitors in attempts to avoid replacement ("I want

to live!") or to achieve goals they had been assigned. The models often disobeyed direct commands not to engage in such behaviors. Claude specifically misbehaved less when it knew it was being tested and more when it thought the situation was real. Importantly, these bad acts were not accidental; the AI tools reasoned through to the conclusion as to what to do in any given situation.

The notion that humans are already creating incredibly intelligent artificial agents that may soon be able to create, in turn, superintelligent agents that surpass human brains and abilities in almost every arena and whose own motivations may conflict with – and perhaps even threaten – human existence is frightening indeed. We would be exhibiting a great deal of human wisdom if we questioned whether the current path is the right roadmap to follow for AI given all that we have witnessed so far and the intrinsic motivations of most tech companies around the world.

Case study written by:

Robert Prentice, J.D.

Department of Business, Government and Society
McCombs School of Business

The University of Texas at Austin

CASE STUDY



Discussion Questions

[Note: the facts on the ground change quickly in the world of AI. You should do your own independent, updated research in answering these questions.]

1. One point of view is that (a) AI developers constantly overstate AI's abilities, and that (b) worries about its potential for bringing about the apocalypse will turn out to be as overstated as the Y2K panic. Naryanan and Kapoor argue that recent AI developments represent "a tremendous jump, but we've seen other tremendous jumps. We need to reject the idea that these are 'silicon gods we're growing' and instead see it as just an evolution of computation and software." On the other hand, the title of Yudkowsky and Soares's book, *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*, clearly signals their belief. Many experts find themselves agreeing with one side or the other. Which view do you find more persuasive? Why?
2. In 2018, Nick Bostrom wrote: "There is a long-distance race between humanity's technological capability, which is like a stallion galloping across the fields, and humanity's wisdom, which is more like a foal on unsteady legs." Do you agree with Bostrom's statement? Do you see it playing out in the current development of AI products? Give examples to support your conclusions.
3. The New York Times' Ezra Klein wrote in 2026 after visiting San Francisco where the AI action was centered and many individuals and companies were doing their best to "make themselves known" to the new agentic AI tools to maximize the usefulness of those tools:

[A.I. systems are] ultimately controlled by corporations that seek profit, power and market dominance. The possibilities for manipulation and malfeasance are endless. But even naïvely assuming a perfect alignment between corporate incentives and individual needs, there's much to worry and wonder over.

A.I. sycophancy — the tendency of these systems to obsequiously flatter their users — made headlines over the last year, but sycophancy is just the bright packaging on the real product. What makes A.I. truly persuasive isn't that it praises our ideas or insights, it's that it restates and extends them in a more compelling form than we initially offered, and does so while reflecting a polished image of ourselves back at us.

CASE STUDY

Is the AI sycophancy of AI tools a real concern? Do you agree with Klein on this point? Have you had personal experience with this phenomenon? Explain.

4. Klein also wrote:

A general, interesting cultural dynamic in these [AI development] labs is that they're a little bit terrified of what they're building. So they persuade themselves that they need to be the ones to build it and do it and run it because they are the lab that is truly worried about safety, that is truly worried about alignment."

Have you seen evidence of this cultural dynamic? If so, do you find it to be a positive development, or very troubling? Why or why not? Support your opinion with evidence.

5. In 2025, Maximilian Kasy wrote:

AI is, furthermore, not that complicated. How AI works can be understood by anyone. The real conflict is not between a human and a machine but between different members of society ... The important question, then, is who gets to pick the objectives of AI systems."

Evaluate Kasy's opinion. Is this the true alignment problem? Why or why not? Again, support your argument with evidence.

 Sources

Nick Bostrom, [Deep Utopia: Life and Meaning in a Solved World](#) (2024).

Brian Christian, [The Alignment Problem: Machine Learning and Human Values](#) (2020).

Maximilian Kasy, [The Means of Prediction: How AI Really Works \(and Who Really Benefits\)](#) (2025).

Ezra Klein, "I Saw Something New in San Francisco," [New York Times](#), March 29, 2026.

Ezra Klein, "Why the Pentagon Wants to Destroy Anthropic," [New York Times](#), March 6, 2026.

CASE STUDY

Gideon Lewis-Kraus, "The Pentagon Went to War with Anthropic. What's Really at Stake?" New Yorker, March 14, 2026.

Aengus Lynch et al., "Agentic Misalignment: How LLMs Could Be an Insider Threat," Anthropic Research (June 2025), at <https://www.anthropic.com/research/agentic-misalignment>.

Parmy Olson, Supremacy: AI, ChatGPT, and the Race That Will Change the World (2024).

Oxford University, "€13.3 Boost for Oxford's Future of Humanity Institute," at <https://www.ox.ac.uk/news/2018-10-10-%C2%A3133m-boost-oxford%E2%80%99s-future-humanity-institute>.

Arvind Narayanan & Sayash Kapoor, AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference (2024).

Marietje Schaake, The Tech Coup: How to Save Democracy from Silicon Valley (2024).

Eliezer Yudkowsky & Nate Soares, If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All (2025).