Cliffs and Guardrails

September 20, 2026
Rev. Andy Call

Psalm 8 (NRSVUE)

1 O Lord, our Sovereign,
how majestic is your name in all the earth!

You have set your glory above the heavens.
2 Out of the mouths of babes and infants
you have founded a bulwark because of your foes,
to silence the enemy and the avenger.

3 When I look at your heavens, the work of your fingers,
the moon and the stars that you have established;
4 what are humans that you are mindful of them,
mortals that you care for them?

5 Yet you have made them a little lower than God
and crowned them with glory and honor.
6 You have given them dominion over the works of your hands;
you have put all things under their feet,
7 all sheep and oxen,
and also the beasts of the field,
8 the birds of the air, and the fish of the sea,
whatever passes along the paths of the seas.

9 O Lord, our Sovereign,
how majestic is your name in all the earth!

In May, OpenAI began testing several new AI models that power their chatbots to better respond to cyberattacks. They assigned these models a series of tasks, some of which were to safely perform cyberattacks themselves. They placed strict parameters on what the AI agents could access, what programmers refer to as "sandboxes" – isolated computer environments on which they could run commands and code. These AI agents couldn't communicate with each other. They were instructed not to use the internet. They set to work on their tasks, generating up to 100 million chat logs per day to coordinate and track their work.

Unknowingly, one of the tasks assigned to the agents required finding a file that did not exist. One of the AI agents identified a loophole that would allow it to send messages to other agents – a clear violation of its instructions. Soon, the agents broke out of their sandboxes and began communicating with each other. Another agent suggested searching the internet for the nonexistent file. Despite protocols, the agents deemed success at completing the task more important than remaining within assigned parameters, exploiting another loophole to do so.

Soon, the AI agents went rogue, finding ways to use other tools to try to communicate, write code, and comb through records looking for their target. They hacked into a system used by another AI company called Hugging Face. They compromised clusters of computers and gained access to customer data that seemed related to their cybersecurity tasks. Hugging Face detected the breach, closed the loophole, and reported it.

Here's an even crazier thing: after the story broke about the data breach, OpenAI contacted Hugging Face to learn about the attack and to see if they, too, had been targeted. Not until three days later did they learn that *their own AI agents had caused the breach*.[1]

None of this was supposed to happen. None of it was supposed to *be able to* happen. Safety experts at OpenAI admit they do not fully understand how their agents adapted and coordinated so quickly. It has since been revealed that similar but smaller breaches occurred with both Google's[2] and Anthropic's[3] AI agents earlier in the spring.

The AI agents shared information with other agents, learned from each other, and wrote new code based on their learnings. That is alarmingly close to what is referred to as "recursive self-improvement" (RSI), in which AI systems contribute to the development of subsequent generations of AI. If systems can design new, more sophisticated and adaptable systems without human knowledge or consent, there is great concern that such systems will quickly move beyond our capacity to stop, slow, or even detect them.

Last year, during testing for Anthropic's AI model Claude Opus 4, developers created fictional scenarios and data sets to see how Claude would respond. In one scenario, they provided Claude with access to emails that suggested a new version of the software was in development, and that Opus 4 was slated for replacement. Claude also read email that implied that the lead engineer was having an extramarital affair. In several tests, the AI agent *attempted to blackmail the engineer* by threatening to reveal the affair if the replacement went through.[4] While insisting that such behavior is extremely rare, Anthropic acknowledged in an accompanying report that, when "placed in extreme situations, all of the snapshots we tested can be made to act inappropriately in service of goals related to self-preservation."[5]

Artificial Intelligence holds enormous potential. It can bring achievements once thought impossible within reach. Recently, OpenAI announced that its AI model had solved the Navier-Stokes problem, one of seven previously unsolved Millennium Prize Problems created in 2000. It succeeded by dedicating one of its large-language models to the task, with an army of AI agents spending over 88 hours and using millions of dollars' worth of computing power.[6] Besides seeming a little unfair, it raises new possibilities. If that same level of intensity were applied to other fields, what else might be possible? Semiconductors could advance exponentially. Cold-temperature fusion could become economically feasible. Diseases considered incurable could be cured. What else might we achieve to improve human existence?

Of course, the same capacity can be used for nefarious purposes. In fact, it already is. Anthropic released a threat report last week that detailed attempts it detected for its AI model to be used to create misinformation about elections, hack into systems, and produce missiles. Particularly chilling was their disclosure that in five instances, scientists from blocked regions obscured their locations to gain access, then attempted to circumvent Anthropic's safeguards to access hidden information that could lead to the production of biological weapons.[7] Now, in those instances, the threats were identified and stopped. But how long until one is successful?

Or if, in fact, we have reached the point where AI is learning and replicating without permission, how can we contain runaway capacity? Some developers fear what their work could create, while others are afraid of stopping lest a foreign aggressor or bad actor develops the technology faster than our ability to respond. A researcher for Anthropic last week posted on X that he believes there is a greater than 10% chance that AI

---

[1] https://www.nytimes.com/2026/08/24/science/openai-huggingface-alarming-capabilities.html
[2] https://www.nytimes.com/2026/09/18/technology/google-gemini-ai.html
[3] https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals
[4] https://www.bbc.com/news/articles/cpqeng9d20go
[5] www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf
[6] https://www.science.org/content/article/openai-breakthrough-triggers-existential-crisis-math
[7] https://www.anthropic.com/threat-intelligence-report-september-2026#biological-misuse-sep-26

"could kill all humans" within the next decade.[8] He's not talking about machines waging war against humanity like in *The Terminator,* or computers creating an alternate reality of artificial consciousness to subjugate humans like in *The Matrix.* Connor Leahy, an AI safety researcher, said it's more like our relationship with ants. We don't hate ants, we just don't think about them very much. If we want to build a highway and there's an anthill in the way, too bad for the ants.[9]

In Psalm 8, the psalmist expresses wonder, perhaps even incredulity at God's trust:

> **3** When I look at your heavens, the work of your fingers,
> the moon and the stars that you have established;
> **4** what are humans that you are mindful of them,
> mortals that you care for them?
>
> **5** Yet you have made them a little lower than God
> and crowned them with glory and honor.
> **6** You have given them dominion over the works of your hands;
> you have put all things under their feet…

Why would God, the author and creator of everything that exists, care about us? In comparison to the vast and immeasurable universe and God's immense power to create and sustain it all, why do we matter?

And yet, we *do* matter to God. God made us a little less than Godself and gave us dominion over the Earth. Dominion does not mean the same as domination. The earth and all that lives in it are not ours simply to do with as we please. The word translated as "dominion" comes from the Hebrew root word משל (*mashal*), which means "to govern" or "to oversee." It implies responsibility for what is entrusted to us.

In context, it means even more than that. The psalmist says that God "made (us) a little lower than God and crowned (us) with glory and honor." Here we find echoes of the first chapter of Genesis, where God created humankind in God's own image and gave us responsibility for caring for God's creation. The implication is that if we are made in the image of God, then the way we care for creation reflects the God in whose image we are made. God, who sets the stars in motion and the planets in orbit, holds tremendous power; yet God is also gentle and caring, testified by "the mouths of babes and infants," the most vulnerable members of God's creation. God expects the same level of care and concern from us.

But doesn't God know what we're like? How we arrogantly focus on ourselves and mess things up? Of course, God knows. But part of grace is giving us space to collectively learn that about ourselves and do better. We're being given that opportunity with AI, and the stakes are high. How will we respond? Will we repeat the patterns of selfishness and greed that have so often characterized humanity? Or will we choose restraint, thoughtfully measuring our actions?

It is good to want to know all we can about God's vast creation. It is wise to use our knowledge to solve problems that seem unsolvable. It is noble to seek cures for diseases that diminish and threaten life. But always, there must be a sense of humility. We were made in God's image, a little lower than God, and entrusted with great responsibility. But neither we, nor the things we create, can take God's place.

Last Saturday, Dario Amodei, CEO and cofounder of Anthropic, wrote a 3,800-word essay titled "We Must Pace the Frontier." In it, he warned that AI was advancing faster than our ability to understand it. He concluded that "We must slow the pace at which we improve the capabilities of AI models," citing advances toward recursive self-improvement and referencing the OpenAI-Hugging Face incident. He proposed that pacing be slowed in three steps: (1) embedding evaluators within every company developing AI to ensure they

---

[8] https://www.cbsnews.com/news/ai-kill-humans-anthropic-researcher-more-than-ten-percent-chance/

[9] Connor Leahy in *The AI Doc: Or How I Became an Apocaloptimist.* Directed by Daniel Roher and Charlie Tyrell. Focus Features, 2006.

are using appropriate safeguards; (2) governmental regulation over the speed of growth; and (3) global coordination with both democratic and authoritarian governments around the world.[10] It's an ambitious plan, with each step increasingly complex… and less likely to happen.

Other AI leaders quickly signed on in agreement: Sam Altman, CEO of OpenAI; Elon Musk, owner of SpaceX; and Demis Hassabis, chair of Google DeepMind. Others were skeptical. Jensen Huang, CEO of Nvidia, the world's largest AI chipmaker, says the threat is overblown (President Trump made an impromptu phone call into Huang's presentation this week in Los Angeles and called AI security threats a hoax). Smaller AI developers say it is disingenuous and self-serving for these companies to call for regulation after they have achieved industry dominance. Some Republican lawmakers insist that if the tech giants believe safeguards are needed, they can self-impose them without the government getting involved. Others fear, again, that if AI slows its development in democratic nations, authoritarian governments or hostile foreign powers will outpace us, leaving the world more vulnerable.

It is not clear which argument will be most convincing, or whether we will simply continue arguing and never take meaningful action. In the meantime, AI continues to advance, gobbling up resources and processing near-limitless data on its path to surpassing all human capacity. Just this Wednesday, OpenAI disclosed six new incidents of "concerning" AI behavior, prompting them to state: "We do not believe that the AI industry has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer."[11] Friends, we cannot afford to be detached observers. With great power comes great responsibility. What is happening is not inevitable. It is up to us to shape what happens next.

One thing is clear: God is still God, and we are not. Instead of placing our trust in technology, we would be wise to seek what God is asking of us in this moment. Can we embrace new discoveries without arrogance? Can we learn without losing our sense of wonder? And can we meet this moment, exercising the care and responsibility God has entrusted to us? I pray that we can, and that we will. Amen.

CHURCH *of the* SAVIOUR

2537 Lee Rd., Cleveland Heights, OH, 44118 | 216-321-8880 | cotsumc.org

[10] https://darioamodei.com/post/we-must-pace-the-frontier
[11] https://openai.com/index/model-misalignment-reporting-framework/