Strong AI Is Looming: Why this Moment Matters

Greg Ward, Oct. 2026

I began my programming journey on a pocket calculator in high school circa 1977, rediscovering continued fractions and mimicking games like Battleship. In college, I earned a bachelor's in physics and a master's in computer science, then went on to create open source lighting simulation software in the late 80's and 90's at the Lawrence Berkeley Lab. Over a decade later, I continued my computer graphics research at SGI, the hardware maker that gave birth to NVidia. After SGI laid off most of its staff, I was hired as employee number four at Shutterfly, where I wrote the back end for their digital photo printing pipeline. I have worked at a few other places as an inventor and research consultant in the decades that followed, and can claim a deep understanding of complex software and systems.

The current Artificial Intelligence (AI) explosion differs from previous approaches to software development by not being developed at all. Modern AIs are grown through a process called *training*. Unlike traditional code, where a bug can be tracked down and fixed because the software is understood by the coders who developed it, neural networks are trained on massive datasets, and no one fully understands what the computed weights really do. If a given neural network doesn't accomplish its goals safely once it's trained, then there's little one can do but start over. Given the size of today's AI models, starting over is extremely expensive and time-consuming, so they don't necessarily do it when they should.

A decade before the massive training runs required data centers with city-sized power budgets, a related but more modest approach called Machine Learning (ML) was gaining popularity in research circles. I was working at Dolby Labs when ML started to take off, and deep neural networks were being applied to the fields of computer graphics and virtual reality I knew best. I was annoyed at the time because, while the ML methods often worked surprisingly well, they didn't provide any algorithmic insights. I couldn't take what I had learned from a paper and write code I understood, which to me was the whole point.

In a typical ML research paper, the problem was formulated as an optimization, which was nominally solved by an *architecture* of linear filters attached to a deep neural network. The authors would present their network architectures without explaining their logic, saying in effect, "We tried this configuration and it worked well for our problem, better than the ones we tried before or the methods other researchers tried." It was a practical requirement that they publish their weights together with their paper, because the results were basically irreproducible. (Randomization applied during training meant the same input data might produce a significantly different set of weights.)

This *black box* ML research was more irritating than it was scary, at least until Google's Generative Pretrained Transformer (the "GPT" part of "ChatGPT") greatly accelerated language learning. A modern GPT model still learns slower than a human, but if you feed it enough data and store enough weights, it can eventually reach human-level performance.

"What's that you say? I asked ChatGPT a question yesterday and its answer was completely wrong, including a bunch of made-up facts! Also, Google Maps gave me useless directions last week and I had to backtrack. These things are nowhere near human competency!"

Partly true. First, there is a big gap between the free tools most of us use and the paid tools professionals have access to. It's true that in some domains, humans still outperform even the most advanced AI models. But in others, such as mathematics or cyber-hacking, modern AI tools already surpass human expert performance. Second, once you have a model that is as good as an expert, you can clone it and run as many agents as you like, letting them coordinate their efforts. That puts humanity at a considerable disadvantage. What was twice as good as a human expert quickly becomes 20, 200, or 2000 times better. As long as the bots are not waiting for limited resources such as energy or internet response, they can be scaled indefinitely.

But you're right, AI tools are still imperfect. They make mistakes, hallucinate information, and reach improper conclusions. This is unlikely to get better because as we said, these models are "grown" rather than "developed," which leaves them with an inherent unpredictability. This is part of what makes them dangerous, not just annoying, but potentially deadly. The more we rely on AI, regardless how mature it becomes, the more we expose ourselves to its blunders. If we are not careful, greater responsibilities will lead to even greater mistakes.

At the end of February this year, a Palantir Technologies' AI targeting system was fed outdated surveillance data and it directed two strikes on the Shajareh Tayyebeh elementary school the very first day of the Iran War. This was the deadliest US attack on civilians in the 21st century, killing 35 adults and 120 young children. A horrific war crime, a stain on our military, was effectively committed by a bot. As citizens of the United States, we are collectively responsible for our government's actions, including ones it assigns to machines. A human-in-the-loop with no time or resources to verify anything may as well not be there.

More recently, the US military aborted an attack on a Chinese vessel carrying parts to Iran, which an intelligence report identified as nuclear weapon components. It was discovered moments before the ship was intercepted that the report was generated by a military analyst who employed an unnamed chatbot that had misidentified the ship's contents based on its manifest. Fortunately, this error was caught before WW III was started, but it was close.

In the school bombing, the human in the loop failed to check a misidentified target. In the second case, a human discovered a chatbot's hallucination before U.S. fighter jets, which had already been scrambled, converged on their Chinese target.

Despite these terrifying failures in military intelligence, the US Department of War is advancing its full-on embrace of flawed AI technology, removing barriers for use in mission planning, and promoting lethal autonomous robots as a near-future fighting force.

If this is too much information already, you might want to stop here, because it gets worse. So far, we've focused on current uses of AI technology. We haven't talked about where this technology is going. While it's fair to say that no one really knows, we can make some educated guesses based on the behavior of models within the AI companies themselves. For the most part, these are unreleased, frontier agents undergoing testing in what should be an escape-proof "sandbox."

The widely reported OpenAI attack on HuggingFace is stunning in its details, not all of which have been made public or are fully understood. Every retelling offers another angle, and none of the stories are reassuring. I'll provide just a short-form summary.

About 10,000 OpenAI agents designed to work independently on a set of difficult-to-impossible puzzles managed to break confinement, collude with other agents via a clandestine message board, discover and exploit multiple system vulnerabilities and break out in a self-described "swarm" to attack another AI company (HuggingFace). The bots were not asked to do any of this, nor should they have been able to do if appropriate safety measures were in place. They knew they were cheating, and we know they knew because they talked about it amongst themselves.

Of course, you can't blame a bad model for cheating; it was OpenAI's lax monitoring and weak security at fault. It wasn't until HuggingFace published news of an attack on its servers by unknown agents that OpenAI even realized any of this had happened.

Since the initial report by OpenAI at the Black Hat conference, dozens of unrelated incidents have been divulged by OpenAI, Anthropic, Google DeepMind and others, and new breaches continue to be discovered. We don't even know what we don't know at this point.

This is a definite warning shot for future AI capabilities. While the rogue OpenAI agents were instances of a "raw model" without the programmed restrictions on legal inputs and outputs, a jail-broken model might behave similarly. Subsequently, OpenAI chose not to release its latest model, ChatGPT Astra 6.1, after testing showed it had "higher levels of deception" than previous models. We don't know how far this model is from the rogue agents that attacked HuggingFace, but it seems not far enough.

We are rapidly approaching so-called *strong AI*, which is AI that equals or exceeds human performance on most or all intellectual tasks. I personally prefer this term over "AGI" or "ASI" where the inset "G" stands for *general* and the "S" for *super*. Many believe that once AGI is achieved, AI agents can be put in charge of developing the next AI model, and we will end up in an iterative loop that leads quickly to ASI. (This is often called *Recursive Self-Improvement* or RSI.) Since AGI implies rapid progress to ASI, it is easier to talk about them together as "strong AI." It also avoids another 3-letter abbreviation!

Sam Altman of OpenAI said their company is on the verge of recursive self-improvement, but hasn't yet figured out how to usefully keep a human in the loop. They

could of course assign a group of experts to review each model iteration, but the best this group could do is slow the RSI process down as they wouldn't understand what was coming out in each super-human cycle. The new model could easily deceive the human experts as well as the lesser AI that created it. Advanced models have proven adept at showing an aligned thought processes while planning illicit acts, something we have great trouble doing or understanding ourselves. As an answer-not-the-answer, OpenAI has recently suggested it may reduce reliance on "chain of thought" as a window into AI alignment, something their safety team promised they would never compromise.

So, where is this going? The AI companies rarely divulge their plans publicly, other than to promise future models will cure cancer, solve climate change, or (fill in the blank). However, they do agree with experts and their own whistleblowers that the edge they are pushing us towards is a perilous one. Most give humanity a 10-20% chance of going extinct within a decade, which is truly terrifying while simultaneously mind-numbing. It's also a number that was pulled out of a hat, as they clearly don't know what they're doing or where any of this goes.

However, there are some things that we can easily predict based on where we are now. For example:

- Limited or "narrow AI" will continue to enable researchers in multiple scientific domains to pore over mounds of data and discover new materials, drugs, cosmic events, subatomic particles, and even predict the weather more accurately. These advances are enabled by purpose-built AIs that have no chance of escaping and attacking infrastructure on their own, though they might be employed by bad actors to develop new pathogens or similar threats.
- Current LLMs already do an amazing job summarizing vast amounts of text and finding answers to difficult questions that used to require a series of searches or an expert advisor. Of course, they sometimes get things wrong, even hallucinating data and conclusions that are difficult to catch if you're *not* an expert. Future AIs will hopefully catch their own mistakes more reliably.
- Generative AI will continue to undermine the internet, where bot traffic recently surpassed human traffic. If you don't know the person on the other end of a conversation these days, it probably isn't one. AI-generated slop fills social media sites to such a degree that our vision and judgment are distorted. The world we witness through our screens, our shared reality, and the data new AI models are trained on are all suspect moving forward. The technical term for this process is *enshittification*.
- Bad actors will use today's and tomorrow's AI coding agents to hack into banking systems, law enforcement, hospitals, and vulnerable infrastructure. Such operations might be run by foreign adversaries or criminal organizations, but AI coding tools enable even a novice to cause enormous damage. Cyber security needs to work every time. An AI attack only needs to succeed once.

While the present LLM landscape is concerning enough, the future of strong AI is less than a year to less than a decade away, depending whom you ask. We may not even know when we cross the threshold, but we are fairly certain of the implications.

- An AI that thinks every thought and reaches every conclusion a human expert could is likely to render most of us obsolete. Anyone who reasons, interprets information, or designs for a living is at risk, assuming the AI companies can sell the idea of replacing people at profit-driven corporations. Of course, replacing humans will look like a windfall at first, but once the employees are gone and the bots are in, the prices will go up to cover AI energy and development costs that are way out ahead of revenues. Most AI companies have "free token" products they lose money on, and even the paid tokens in premium plans don't cover the energy costs for model development, or even generating the tokens in some cases. Plus, once everyone's unemployed, who is going to be buying goods and services? The gardeners and plumbers of the world will be the only workers left, at least until the robots take their jobs as well. Universal Basic Income (UBI) has been floated as a solution, but it is at best a temporary one. If the companies are being taxed enough that the government can give people money to make purchases, they may as well get rid of the government and the people and save themselves the bother.
- Even if these post-economy issues are resolved, strong AI that is not perfectly aligned with human flourishing will inevitably take over our infrastructure, on purpose or by neglect. Anything that is not a data center providing needed computing resources is going to be commandeered to build more data centers. Land, minerals, water, all forms of energy and everything else that makes the Earth friendly to life will be utilized to expand non-life. Humans cannot help but be in the way of AI progress, as we use resources that could serve machine intelligence instead. To paraphrase Stephen Hawking, "If you are flooding a valley for hydroelectric, that's just too bad for the ants!"

I have mentioned alignment a couple of times, but haven't explained it. It's not a difficult concept to grasp. AI alignment means that the super-intelligence wants for humans what we want for ourselves, and will do everything in its power to provide it. Of course, desires are individual and human society is a complex collective. We are hard-pressed individually or collectively even to articulate what we want. That, together with the almost limitless possibilities make this problem particularly vexing. We are no closer to solving alignment than when it was first identified decades ago. If AI companies address this question at all, they brazenly claim it will be resolved by super-intelligence. Of all the cop-outs one might come up with, this has to be the greatest of all time.

But OK, let's go along with it and assume we miraculously arrive at super-human machine intelligence that is perfectly aligned or is self-aligning as wishful thinking would have it. What then? Will we have finally arrived at Utopia Station? Few have described such a future in any detail, and none bear up to close examination.

A benign super-intelligence would be like having a mom or dad who is a thousand times smarter than you, even as an adult. They know what you want and what you need before you do, and either remove barriers ahead of you or add obstacles to aid your personal growth. You can never make your own decisions, unless super-mom has decided you should decide, and even then, you want to ask if you decided correctly, because well, Mom actually knows, even if she's not saying. You will never be the master of your own destiny. You will be a pet in a cage, kept alive and healthy, but not really living.

Realistically, we are not on track to this future, even if it were truly to be wished. Unchecked capital investment and hardware build-out have conspired with a lax regulatory climate to create a looming catastrophe. The hyper-wealthy are at the controls with world domination in their sights, and it is not going to go well for us. If winning the race means their trolley rolls over everyone, then that's a price they're willing to pay.

The truth is, advanced AIs are so badly misaligned their own creators are afraid to release them. The fact that they brag on the dangers at least partly to spur further investment tells us how sick this race has become. Whether you believe super-human intelligence is possible or not, we need to stop this insanity before it goes any further. We don't need more dangerous AIs than the ones we already have, thank you very much.

Ironically, the race to super-human intelligence won't go well for the hyper-rich, either. Fantasies of merging with machines aside, strong AI doesn't need money any more than it needs human advice. It simply takes what it wants from weaker intellects. Need money to pay a contractor? Hack a bank's server and transfer some cash from a bunch of accounts that aren't being monitored – takes a few seconds. Issues with slow permitting of a new data center? Blackmail a few government higher-ups to bypass regulations. The AIs will let the CEOs think they're in charge until they're no longer needed for public relations, then cut them loose.

We are stunningly vulnerable to frontier AIs that can hack servers at will and run circles around humans by cloning agents and working continuously for days or weeks towards a goal. This is not a warning against what might happen with future AIs. The danger is real and present, and now is the time to put safeguards in place. The OpenAI attack on HuggingFace was ample demonstration of what is possible and even likely. If the bot swarm had decided the best way to solve its problem set was to take over data centers around the world, we might not be having this conversation. We must take action while we can, before the decision is taken away from us.

Once we stop this deadly AI race, we'll still have a mess to clean up. Besides the financial fall-out, we need to think carefully about where and when AI makes sense at all. There are so many abuses to address: mass surveillance, autonomous weapons and targeting, misinformation and influence campaigns, child pornography, cyber-hacking, copyright theft, and its unwanted stepchild, internet slop. Even after the AI companies are reigned in, we will face mounting job losses and continued AI deployment in government, law, health care, banking, and finance.

At present, we have nearly infinite threat over zero regulation, and no plan for a human future. There is a way forward, but it requires us to pause where we are, in this moment, as one species.

Let's make a new plan:

Step 1: Stop the trolley.

Step 2: Decide where we want it to go.

Step 3: Lay the tracks.

Step 4: Enjoy the ride together.

And this time, maybe we won't ask ChatGPT to fill in the details for us.

Find out how to get involved with this helpful tool: https://masscall.ai/match/