Preventing AI from Wiping Out Humanity

By Joe Weber

The debate over existential AI risk is focused mainly on prevention: government regulation, voluntary restraint by AI companies, testing, safeguards, and international agreements. All are necessary. But taken together they will still be insufficient.

The United States cannot simply slow advanced AI development while China races ahead. And even if the major powers eventually cooperate, AI is not like nuclear weapons. Nuclear weapons require factories, centrifuges, and fissile material. Dangerous AI may eventually require little more than computing power, code, and a connection. Once powerful models and techniques exist, they can be copied, stolen, modified, and deployed by bad actors almost anywhere.

So we need to ask a harder question: What happens when prevention fails? We have spent enormous time and energy asking how to keep AI from becoming dangerous. We have spent far less asking how we would stop it once it is.

A sufficiently powerful AI in the hands of a hostile government, terrorist organization, or criminal group could potentially attack cyber networks, energy systems, communications, financial infrastructure, autonomous weapons, and/or biological systems. And someday the threat may not come only from humans using AI. It could come from an advanced AI acting autonomously, pursuing objectives that conflict with ours.

That possibility demands a shift in how we think about AI safety. We do not rely on diplomacy alone to prevent missile attacks. We build radar, interceptors, and command systems capable of detecting and stopping them. AI may require the same combination of prevention and active defense.

Humanity should acquire an AI immune system. Just as the biological immune system detects harmful pathogens, distinguishes them from healthy cells, and mobilizes defenses to neutralize them, we may need defensive AI able to recognize hostile AI, contain it, and stop it before it spreads.

Regulation is the guardrail. Defense is the airbag.

We should begin building AI systems specifically designed to detect hostile machine activity, protect critical infrastructure, identify anomalous behavior, respond at machine speed, and help humans isolate dangerous systems before they can do widespread damage. And defense may not be enough.

If a rogue AI were continuously probing networks, discovering vulnerabilities, replicating itself, and moving from system to system, simply blocking individual attacks could become impossible. We may also need tightly controlled capabilities to identify the source, cut off its access to computing and communications, and disable it. Humanity may eventually need not only AI defense, but the ability to neutralize hostile AI.

That does not mean unleashing autonomous systems to fight one another without human control. It means preparing now, under strict human authority, for a possibility we hope never occurs. There is, of course, a profound irony here: Defensive AI must be powerful enough to defeat hostile AI without itself becoming another uncontrolled actor. That may become one of the central engineering challenges of the AI age.

The United States should therefore launch a major national AI-defense initiative involving leading AI companies, cybersecurity firms, biotechnology experts, critical-infrastructure operators, universities, national-security agencies, and counterparts in allied nations. Its mission could be stated in one sentence: If an extremely capable AI – or someone wielding one – attacks civilization, how do we stop it?

We should be running carefully contained AI war games: offensive systems probing simulated power grids, communications networks, financial systems, and biological defenses while defensive AI learns to detect and defeat them. Such exercises would themselves be hazardous and should be conducted in hardened, physically and digitally isolated environments, with multiple containment layers and no pathway to real critical infrastructure. But we should still conduct AI war games before we have an AI war.

The decisive AI arms race may not ultimately be America versus China. It may be hostile AI versus defensive AI – with humanity's security depending on which side is better prepared.

We should continue developing increasingly powerful artificial intelligence. Its benefits will be extraordinary. But we should also recognize a hard truth: Someday AI could become an adversary.

The goal cannot merely be to make every AI safe. The goal must be to make humanity safe even when some AI is not.

Joe Weber
joeweber@alum.mit.edu
617-275-6260
https://www.linkedin.com/in/joe-weber-ms-mba-327b9a6/