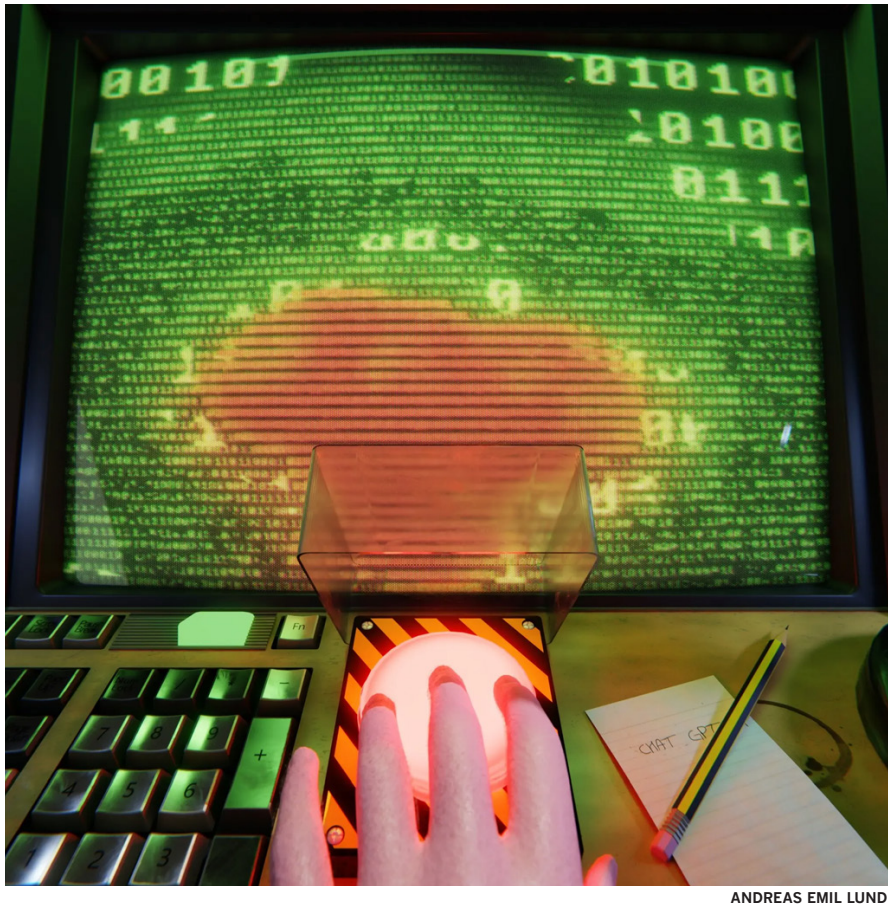
The New York Times
NYTIMES.COM
"All the News That's Fit to Print" TUESDAY, JULY 25, 2023 Reprinted With Permission

Opinion

GUEST ESSAY

Our Oppenheimer Moment: The Creation of A.I. Weapons

By ALEXANDER C. KARP



ANDREAS EMIL LUND

In 1942, J. Robert Oppenheimer, the son of a painter and a textile importer, was appointed to lead Project Y, the military effort established by the Manhattan Project to develop nuclear weapons. Oppenheimer and his colleagues worked in secret at a remote laboratory in New Mexico to discover methods for purifying uranium and ultimately to design and build working atomic bombs.

He had a bias toward action and inquiry. "When you see something that is technically sweet, you go ahead and do it," he told a government panel that would later assess his fitness to remain privy to U.S. secrets. "And you argue about what to do about it only after you have had your technical success. That is the way it was with the atomic bomb." His security clearance was revoked shortly after his testimony, effectively ending his career in public service.

Oppenheimer's feelings about his role in conjuring the most destructive weapon of the age would shift after the bombings of Hiroshima and Nagasaki. At a lecture at the Massachusetts Institute of Technology in 1947, he observed that the physicists involved in the development of the bomb "have known sin" and that this is "a knowledge which they cannot lose."

We have now arrived at a similar crossroads in the science of computing, a crossroads that connects engineering and ethics, where we will again have to choose whether to proceed with the development of a technology whose power and potential we do not yet fully apprehend.

The choice we face is whether to rein in or even halt the development of the most advanced forms of artificial intelligence, which some argue may threaten or someday supersede humanity, or to allow more unfettered experimentation with a technology that has the potential to shape the international politics of this century in the way nuclear arms shaped the last one.

The emergent properties of the latest large language models — their ability to stitch together what seems to pass for a primitive form of knowledge of the workings of our world — are not well understood. In the absence of understanding, the collective reaction to early encounters with this novel technology has been marked by an uneasy blend of wonder and fear.

Some of the latest models have a trillion or more parameters, tunable variables within a computer algorithm, representing a scale of processing that is impossible for the human mind to begin to comprehend. We have learned that the more

parameters a model has, the more expressive its representation of the world and the richer its ability to mirror it.

What has emerged from that trillion-dimensional space is opaque and mysterious. It is not at all clear — not even to the scientists and programmers who build them — how or why the generative language and image models work. And the most advanced versions of the models have now started to demonstrate what one group of researchers has called "sparks of artificial general intelligence," or forms of reasoning that appear to approximate the way that humans think.

In one experiment that tested the capabilities of GPT-4, the language model was asked how one could stack a book, nine eggs, a laptop, a bottle and a nail "onto each other in a stable manner." Attempts at prodding more primitive versions of the model into describing a workable solution to the challenge had failed.

GPT-4 excelled. The computer explained that one could "arrange the nine eggs in a three-by-three square on top of the book, leaving some space between them," and then "place the laptop on top of the eggs," with the bottle going on top of the laptop and the nail on top of the bottle cap, "with the pointy end facing up and the flat end facing down."

It was a stunning feat of "common sense," in the words of Sébastien Bubeck, the French lead author of the study who taught computer science at Princeton University and now works at Microsoft Research.

It is not just our own lack of understanding of the internal mechanisms of these technologies but also their marked improvement in mastering our world that has inspired fear. A growing group of leading technologists has issued calls for caution and debate before pursuing further technical advances. An open letter to the engineering community calling for a six-month pause in developing more advanced forms of A.I. has received more than 33,000 signatures. On Friday, at a White House meeting with President Biden, seven companies that are developing A.I. announced their commitment to a set of broad principles intended to manage the risks of artificial intelligence.

In March, one commentator published an essay in Time magazine arguing that "if somebody builds a too-powerful A.I., under present conditions," he expects "that every single member of the human species and all biological life on Earth dies shortly thereafter."

Concerns such as these regarding the further development of artificial intelligence are not unjustified. The software that we are building can enable the deployment of lethal weapons. The potential integration of weapons systems with increasingly autonomous artificial intelligence software necessarily brings risks.

But the suggestion to halt the development of these technologies is misguided.

Some of the attempts to rein in the advance of large language models may be driven by a distrust of the public and its ability to appropriately weigh the risks and rewards of the technology. We should be skeptical when the elites of Silicon Valley, who for years recoiled at the suggestion that software was anything but our salvation as a species, now tell us that we must pause vital research that has the potential to revolutionize everything from military operations to medicine.

A significant amount of attention has also been directed at the policing of language that chatbots use and to patrolling the limits of acceptable discourse with the machine. The desire to shape these models in our image, and to require them to conform to a particular set of norms governing interpersonal interaction, is understandable but may be a distraction from the more fundamental risks that these new technologies present. The focus on the propriety of the speech produced by language models may reveal more about our own preoccupations and fragilities as a culture than it does the technology itself.

Our attention should instead be more urgently directed at building the technical architecture and regulatory framework that would construct moats and guardrails around A.I. programs' ability to autonomously integrate with other systems, such as electrical grids, defense and intelligence networks, and our air traffic control infrastructure. If these technologies are to exist alongside us over the long term, it will also be essential to rapidly construct systems that allow more seamless collaboration between human operators and their algorithmic counterparts, to ensure that the machine remains subordinate to its creator.

We must not, however, shy away from building sharp tools for fear they may be turned against us.

A reluctance to grapple with the often grim reality of an ongoing geopolitical struggle for power poses its own danger. Our adversaries will not pause to indulge in theatrical debates about the merits of developing technologies with critical military and national security applications. They will proceed.

This is an arms race of a different kind, and it has begun.

Our hesitation, perceived or otherwise, to move forward with military applications of artificial intelligence will be punished. The ability to develop the tools required to deploy force against an opponent, combined with a credible threat to use such force, is often the foundation of any effective negotiation with an adversary.

The underlying cause of our cultural hesitation to openly pursue technical superiority may be our collective sense that we have already won. But the certainty with which many believed that history had come to an end, and that Western liberal democracy had emerged in permanent victory after the struggles of the 20th century, is as dangerous as it is pervasive.

We must not grow complacent.

The ability of free and democratic societies to prevail requires something more than moral appeal. It requires hard power, and hard power in this century will be built on software.

Thomas Schelling, an American game theorist who taught economics at Harvard and Yale, understood the relationship between technical advances in the development of weaponry and the ability of such weaponry to shape political outcomes.

"To be coercive, violence has to be anticipated," he wrote in the 1960s as the United States grappled with its military escalation in Vietnam. "The power to hurt is bargaining power. To exploit it is diplomacy — vicious diplomacy, but diplomacy."

While other countries press forward, many Silicon Valley engineers remain opposed to working on software projects that may have offensive military applications, including machine learning systems that make possible the more systematic targeting and elimination of enemies on the battlefield. Many of these engineers will build algorithms that optimize the placement of ads on social media platforms, but they will not build software for the U.S. Marines.

In 2019, Microsoft faced internal opposition to accepting a defense contract with the U.S. Army. "We did not sign up to develop weapons," employees wrote in an open letter to corporate management.

A year earlier, an employee protest at Google preceded the company's decision not to renew a contract for work with the U.S. Department of Defense on a critical system for planning and executing special forces operations around the world. "Building this technology to assist the U.S. government in military surveillance — and potentially lethal outcomes — is not acceptable," Google employees wrote in an open letter to Sundar Pichai, the company's chief executive officer.

I fear that the views of a generation of engineers in Silicon Valley have meaningfully drifted from the center of gravity of American public opinion. The preoccupations and political instincts of coastal elites may be essential to maintaining their sense of self and cultural superiority but do little to advance the interests of our republic. The wunderkinder of Silicon Valley — their fortunes, business empires and, more fundamentally, their entire sense of self — exist because of the nation that in many cases made their rise possible. They charge themselves with constructing vast technical empires but decline to offer support to the state whose protections and underlying social fabric

have provided the necessary conditions for their ascent. They would do well to understand that debt, even if it remains unpaid.

Our experiment in self-government is fragile. The United States is far from perfect. But it is easy to forget how much more opportunity exists in this country for those who are not hereditary elites than in any other nation on the planet.

Our company, Palantir Technologies, has a stake in this debate. The software platforms that we have built are used by U.S. and allied defense and intelligence agencies for functions like target selection, mission planning and satellite reconnaissance. The ability of software to facilitate the elimination of an enemy is a precondition for its value to the defense and intelligence agencies with which we work. At Palantir, we are fortunate that our interests as a company and those of the country in which we are based are fundamentally aligned. In the wake of the invasion of Ukraine, for example, we were often asked when we decided to pull out of Russia. The answer is never, because we were never there.

A more intimate collaboration between the state and the technology sector, and a closer alignment of vision between the two, will be required if the United States and its allies are to maintain an advantage that will constrain our adversaries over the long term. The preconditions for a durable peace often come only from a credible threat of war.

In the summer of 1939, from a cottage on the North Fork of Long Island, Albert Einstein sent a letter — which he had worked on with Leo Szilard and others — to President Franklin Roosevelt, urging him to explore building a nuclear weapon, and quickly. The rapid technical advances in the development of a potential atomic weapon, Einstein and Szilard wrote, "seem to call for watchfulness and, if necessary, quick action on the part of the administration," as well as a sustained partnership founded on "permanent contact maintained between the administration" and physicists.

It was the raw power and strategic potential of the bomb that prompted their call to action then. It is the far less visible but equally significant capabilities of these newest artificial intelligence technologies that should prompt swift action now.

Alexander Karp is the C.E.O. of Palantir Technologies, a company that creates data analysis software and works with the U.S. Department of Defense.