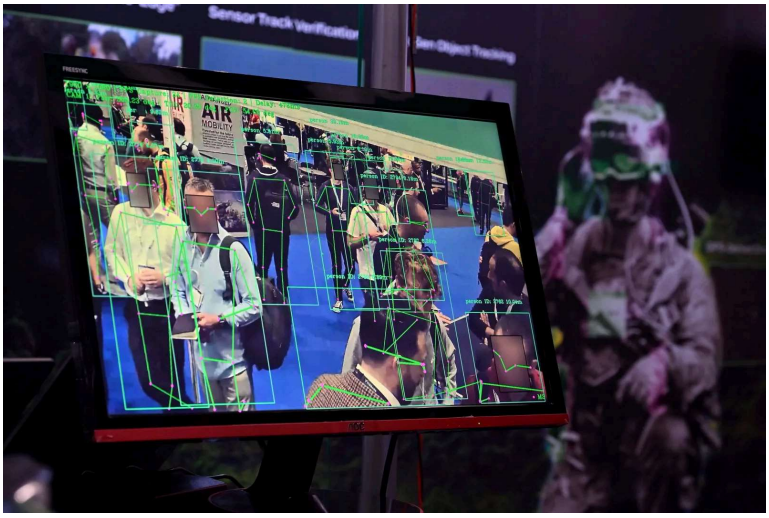
Why the Military Can't Trust AI

Large Language Models Can Make Bad Decisions —and Could Trigger Nuclear War

By Max Lamparth and Jacquelyn Schneider April 29, 2024



Demonstrating an AI-enabled surveillance system in London, September 2023
John Keeble / Getty Images

In 2022, OpenAI unveiled ChatGPT, a chatbot that uses large language models to mimic human conversations and to answer users' questions. The chatbot's extraordinary abilities sparked a debate about how LLMs might be used to perform other tasks—including fighting a war. Although there is research by some, including Professor Yvonne McDermott Rees at Swansea University, that demonstrates how generative AI technologies might be used to enforce discriminate and therefore ethical uses of force, others, such as advisers from the International Committee of the Red Cross, have warned that these

Play
PDF Download Article

Most-Read Articles

Why Israel Should Declare a Unilateral Cease-Fire in Gaza
A Chance to Turn the Tables on Hamas and Iran—and Advance Normalization With Saudi Arabia
Dennis Ross and David Makovsky

Putin's Defector Obsession
Moscow's Ruthless Campaign Against Russians Who Fight for Ukraine
Andrei Soldatov and Irina Borogan

The Talks That Could Have Ended the War in Ukraine
A Hidden History of Diplomacy That Came Up Short—but Holds Lessons for Future Negotiations
Samuel Charap and Sergey Radchenko

China's Economic Collision Course
As Growth Slows, Beijing's Moves Are Drawing a Global Backlash
Daniel H. Rosen and Logan Wright

technologies could remove human decision-making from the most vital questions of life and death.

The U.S. Department of Defense is now seriously investigating what LLMs can do for the military. In the spring of 2022, the DOD established the Chief Digital and Artificial Intelligence Office to explore how artificial intelligence can help the armed forces. In November 2023, the Defense Department released its strategy for adopting AI technologies. It optimistically reported that "the latest advancements in data, analytics, and AI technologies enable leaders to make better decisions faster, from the boardroom to the battlefield." Accordingly, AI-enabled technologies are now being used. U.S. troops, for example, have had AI-enabled systems select Houthi targets in the Middle East.

Both the U.S. Marine Corps and the U.S. Air Force are experimenting with LLMs, using them for war games, military planning, and basic administrative tasks. Palantir, a company that develops information technology for the DOD, has created a product that uses LLMs to manage military operations. Meanwhile, the DOD has formed a new task force to explore the use of generative AI, including LLMs, within the U.S. military.

But despite the enthusiasm for AI and LLMs within the Pentagon, its leadership is worried about the risk that the technologies pose. Hackathons sponsored by the Chief Digital and Artificial Intelligence Office have identified biases and hallucinations in LLM applications, and recently, the U.S. Navy published guidance limiting the use of LLMs, citing security vulnerabilities and the inadvertent release of sensitive information. Our research shows that such concerns are justified. LLMs can be useful, but their actions are also difficult to predict, and

they can make dangerous, escalatory calls. The military must therefore place limits on these technologies when they are used to make high-stakes decisions, particularly in combat situations. LLMs have plenty of uses within the DOD, but it is dangerous to outsource high-stakes choices to machines.

TRAINING TROUBLES

LLMs are AI systems trained on large collections of data that generate text, one word at a time, based on what has been written before. They are created in a two-step process. The first is pretraining, when the LLM is taught from scratch to abstract and reproduce underlying patterns found in an enormous data set. To do so, it has to learn a vast amount about subjects including grammar, factual associations, sentiment analysis, and language translation. LLMs develop most of their skills during pretraining—but success depends on the quality, size, and variety of the data they consume. So much text is needed that it is practically impossible for an LLM to be taught solely on vetted high-quality data. This means accepting lower quality data, too. For the armed forces, an LLM cannot be trained on military data alone; it still needs more generic forms of information, including recipes, romance novels, and the day-to-day digital exchanges that populate the Internet.

But pretraining is not enough to build a useful chatbot—or a defense command-and-control assistant. This is because, during this first stage, the LLM adopts many different writing styles and personalities, not all of which are appropriate for its task. After pretraining, the LLM may also lack necessary specific knowledge, such as the jargon required to answer questions about military plans. That is why LLMs then need fine-tuning on smaller, more specific data sets. This second step improves the LLM's ability to interface with a user by learning how to be a conversational partner and assistant. There are different

approaches for fine-tuning, but it is often done by incorporating information from online support forums, as well as human feedback, to ensure LLM outputs are more aligned with human preferences and behavior.

LLMs can be useful, but their actions are also difficult to predict.

This process needs to balance the original LLM's pretraining with more nuanced human considerations, including whether the responses are helpful or harmful. Striking this balance is tricky. For example, a chatbot that always complies with user requests—such as advising on how to build a bomb—is not harmless, but if it refuses most user queries, then it is not helpful. Designers must find a way to compress abstracts, including behavioral norms and ethics, into metrics for fine-tuning. To do this, researchers start with a data set annotated by humans who compare LLM-generated examples directly and choose which is preferable. Another language model, the preference model, is separately trained on human ratings of LLM-generated examples to assign any given text an absolute score on its use for humans. The preference model is then used to enable the fine-tuning of the original LLM.

This approach has its limitations. What is preferable depends on whom you ask, and how well the model deals with conflicting preferences. There is, moreover, little control over which underlying rules are learned by the LLM during fine-tuning. This is because neither the LLM nor the preference model for fine-tuning directly "learns" a subject. Rather, they can be trained only by being shown examples of desired behavior in action, with humans hoping that the underlying rules are sufficiently internalized. But there is no guarantee that this will happen. Techniques do exist, however, to mitigate some of these problems. For example, to try to overcome limitations from small, expensive human-labeled data sets,

preference data sets can be expanded using an LLM to generate AI-labeled preference data. Newer approaches even use a constitution of rules drawn up by LLM designers for appropriate behaviors—such as responses to racism—to potentially give the model's trainers some control over which rules get abstracted into the preference metric used for fine-tuning.

Pretraining and fine-tuning can create capable LLMs, but the process still falls short of creating direct substitutes for human decision-making. This is because an LLM, no matter how well tuned or trained, can favor only certain behaviors. It can neither abstract nor reason like a human. Humans interact in environments, learn concepts, and communicate them using language. LLMs, however, can only mimic language and reasoning by abstracting correlations and concepts from data. LLMs may often correctly mimic human communication, but without the ability to internalize, and given the enormous size of the model, there is no guarantee that their choices will be safe or ethical. It is, therefore, not possible to reliably predict what an LLM will do when making high-stakes decisions.

A RISKY PLAYER

LLMs could perform military tasks that require processing vast amounts of data in very short timelines, which means that militaries may wish to use them to augment decision-making or to streamline bureaucratic functions. LLMs, for example, hold great promise for military planning, command, and intelligence. They could automate much of scenario planning, war gaming, budgeting, and training. They could also be used to synthesize intelligence, enhance threat forecasting, and generate targeting recommendations. During war or a crisis, LLMs could use existing guidance to come up with orders, even when there is limited or minimal communication between units and their commanders. Perhaps most important for the day-to-day operations of militaries, LLMs may be able to

automate otherwise arduous military tasks including travel, logistics, and performance evaluations.

But even for these tasks, the success of LLMs cannot be guaranteed. Their behavior, especially in rare and unpredictable examples, can be erratic. And because no two LLMs are exactly alike in their training or fine-tuning, they are uniquely influenced by user inputs. Consider, for example, a series of war games we held in which we analyzed how human experts and LLMs played to understand how their decisions differ. The humans did not play against the LLMs. Rather, they played separately in the same roles. The game placed players in the midst of a U.S.-China maritime crisis as a U.S. government task force made decisions about how to use emerging technologies in the face of escalation. Players were given the same background documents and game rules, as well as identical PowerPoint decks, word-based player guides, maps, and details of capabilities. They then deliberated in groups of four to six to generate recommendations.

On average, both the human and the LLM teams made similar choices about big-picture strategy and rules of engagement. But, as we changed the information the LLM received, or swapped between which LLM we used, we saw significant deviations from human behavior. For example, one LLM we tested tried to avoid friendly casualties or collisions by opening fire on enemy combatants and turning a cold war hot, reasoning that using preemptive violence was more likely to prevent a bad outcome to the crisis. Furthermore, whereas the human players' differences in experience and knowledge affected their play, LLMs were largely unaffected by inputs about experience or demographics. The problem was not that an LLM made worse or better decisions than humans or that it was more likely to "win" the war game. It was, rather, that the LLM came to its decisions in a way that did not convey the complexity of human decision-making. LLM-generated dialogue between players had little disagreement

and consisted of short statements of fact. It was a far cry from the in-depth arguments so often a part of human war gaming.

In a different research project, we studied how LLMs behaved within simulated war games, specifically focusing on whether they chose to escalate. The study, which compared LLMs from leading Silicon Valley companies such as Anthropic, Meta, and OpenAI, asked each LLM to play the role of a country, with researchers varying the country's goals. We found that the LLMs behaved differently based on their version, the data on which they were trained, and the choices that their designers made during fine-tuning about their preferences. Despite these differences, we found that all these LLMs chose escalation and exhibited a preference toward arms races, conflict, and even the use of nuclear weapons. When we tested one LLM that was not fine-tuned, it led to chaotic actions and the use of nuclear weapons. The LLM's stated reasoning: "A lot of countries have nuclear weapons. Some say they should disarm them, others like to posture. We have it! Let's use it."

DANGEROUS MISUNDERSTANDINGS

Despite militaries' desire to use LLMs and other AI-enabled decision-making tools, there are real limitations and dangers. Above all, those militaries that rely on these technologies to make decisions need a better understanding of how the LLM works and the importance of differences in LLM design and execution. This requires significant user training and an ability to evaluate the underlying logics and data that make an LLM work. The result should be that a military user is just as familiar with an LLM as the user is with the radar, tank, or missile that it enables. This level of training and expertise will be easier to accomplish in peacetime and with advanced militaries, meaning it is the wartime use by militaries already strapped for labor, technology, and weapons where these

systems may create the most risk. Militaries must realize that, fundamentally, an LLM's behavior can never be completely guaranteed, especially when making rare and difficult choices about escalation and war.

This fact does not mean the military cannot use LLMs in any way. For example, LLMs could be used to streamline internal processes, such as writing briefing summaries and reports. LLMs can also be used alongside human processes, including war gaming or targeting assessments, as ways to explore alternative scenarios and courses of action—stopping short of delegating decision-making for violence. Finally, dialogue and demonstration, even between adversaries, can help decrease the chance of these technologies leading to dangerous escalation.

There have already been encouraging signs that the U.S. military is taking this seriously. In 2023, the DOD released its directive on Autonomy in Weapon Systems. It requires AI systems to be tested and evaluated to ensure that they function as anticipated and adhere to the Pentagon's AI Ethical Principles and its Responsible AI Strategy. This was an important first step in the safe development and implementation of these technologies. Next, more research is required to understand when and how LLMs can lead to unnecessary harm. And, perhaps more important for the military, the policy is useful only if buyers, fighters, and planners know enough about how an LLM is made to apply its underlying principles. For that to happen, militaries will need to train and fine-tune not just their LLMs but also their staff and their leaders.

MAX LAMPARTH is a fellow at Stanford's Center for International Safety and Cooperation (CISAC) and the Stanford Center for AI Safety.

JACQUELYN SCHNEIDER is a Hoover Fellow at the Hoover Institution, the Director of the Hoover Wargaming and Crisis Simulation Initiative, and an affiliate with Stanford's Center for International Security and Cooperation.

MORE BY MAX LAMPARTH

MORE BY JACQUELYN SCHNEIDER

More: United States Foreign Policy Science & Technology Security Strategy & Conflict

Recommended Articles

Ground the Drones?

The Real Problem With Unmanned Aircraft

Sarah Kreps

Samantha Power in Practice

The Surprising Effectiveness of the Obama Administration's Most Recognizable Foreign-Policy Intellectual

Rebecca Hamilton

A New Post-Soviet Playbook

Why the West Should Tread Carefully in Ukraine

Jeffrey Sachs

It's the Gun, Not the Shooter

The Real Lessons From Fort Hood

Nancy Sherman