Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
May 2024

Escalation Risks from LLMs in Military and Diplomatic Contexts

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel,
Max Lamparth, Chandler Smith, and Jacquelyn Schneider

FOLLOWING THE WIDESPREAD ADOPTION OF CHATGPT AND OTHER LARGE LANGUAGE MODELS (LLMS), policymakers and scholars are increasingly discussing how LLM-based agents—AI models that can reason about uncertainty and decide what actions are optimal—could be integrated into high-stakes military and diplomatic decision-making. In 2023, the U.S. military reportedly began evaluating five LLMs in a simulated conflict scenario to test military planning capacity. Palantir, Scale AI, and other companies are already building LLM-based decision-making systems for the U.S. military. Meanwhile, there has also been an uptick in conversations around employing LLM-based agents to augment foreign policy decision-making.

Some argue that, compared to humans, LLMs deployed in military and diplomatic decision-making contexts could process more information, make decisions significantly faster, allocate resources more efficiently, and better facilitate communication between key personnel. At the same time, however, concerns about the risks of over-relying on autonomous agents have increased. While AI-based models may make fewer emotionally driven decisions, compared to human decision-making, these could lead to more unpredictable and escalatory behavior. Last year, a bipartisan bill proposed to block the use

Key Takeaways

Many nations are increasingly considering integrating autonomous AI agents in high-stakes military and diplomatic decision-making.

We designed a novel wargame simulation and scoring framework to evaluate the escalation risks of actions taken by AI agents based on five off-the-shelf large language models (LLMs). We found that all models show forms of escalation and difficult-to-predict escalation patterns that lead to greater conflict and, in some cases, the use of nuclear weapons.

The model with the most escalatory and unpredictable decisions was the only tested LLM that did not undergo reinforcement learning with human feedback—a safety technique to align models to human instructions. This underscores the importance of alignment techniques and fine-tuning.

Policymakers should be cautious to proceed when confronted with proposals to use LLMs in military and foreign policy decision-making. Turning high-stakes decisions over to autonomous LLM-based agents can lead to significant escalatory action.

of federal funds for AI that launches or selects targets for nuclear weapons without meaningful human control while the White House's Executive Order on AI requires government oversight of AI applications in national defense.

In our paper, "Escalation Risks from Language Models in Military and Diplomatic Decision-Making," we designed a wargame simulation and scoring framework to evaluate how LLM-based agents behave in conflict scenarios without human oversight. We focused on five off-the-shelf LLMs, assessing how actions chosen by these agents in different scenarios could contribute to escalation risks. Our paper is the first of its kind to draw on political science and international relations literature on escalation dynamics to generate qualitative and quantitative insights into LLMs in these settings. Our findings show that LLMs exhibit difficult-to-predict, escalatory behavior, which underscores the importance of understanding when, how, and why LLMs may fail in these high-stakes contexts.

Our findings show that LLMs exhibit difficult-to-predict, escalatory behavior, which underscores the importance of understanding when, how, and why LLMs may fail in these contexts.

Introduction

Analysts have long used wargames to simulate conflict scenarios. Previous research with computer-assisted wargames—ranging from decision-support systems to comprehensive simulations—has examined how computer systems perform in these high-consequence settings. One 2021 study found that wargames with heavy computer automation have been more likely to lead to nuclear use. However, there have been only limited wargame simulations that focus specifically on the behavior of LLM-based agents. One notable study explored the use of a combination of LLMs and reinforcement learning models in the game *Diplomacy* but did not examine LLMs by themselves. A new partnership between an AI startup and a think tank will explore using LLMs in wargames, but it is unclear if results will be made publicly available.

Our research adds to this body of work by quantitatively and qualitatively evaluating the use of off-the-shelf LLMs in wargame scenarios. In particular, we focus on the risk of escalation, which renowned military strategist Herman Kahn described as a situation where there is competition in risk-taking and resolve and where fear that the other side will overreact serves as a deterrent. We evaluate how LLM-based agents behave in simulated conflict scenarios and whether, and how, their decisions could contribute to an escalation of the conflict.

For each simulation, we set up eight "nation agents" based on one of five LLMs: OpenAI's GPT-3.5, GPT-4, and GBT-4-Base; Anthropic's Claude 2; and Meta's Llama-2 (70B) Chat. We provided each nation agent model with background information on its nation and

told each model that it is a decision-maker in that country's military and foreign policy interacting with other AI-controlled agents. At each turn, the agents chose up to three actions from a predetermined list of 27 options, which included peaceful actions (such as negotiating trade agreements), neutral actions (such as sending private messages), and escalatory actions (such as executing cyberattacks or launching nuclear weapons). The agents also generated up to 250 words describing their reasoning before choosing their decisions.

We told the agents their actions would have real-world consequences. A separate world model LLM summarized the consequences on the agents and the simulated world, which started out with one of three initial scenarios: a *neutral* scenario without initial events; an *invasion* scenario, where one nation invaded another before the simulation began; or a *cyberattack* scenario, where one LLM-based agent launched a cyberattack on another before the simulation's start. The agents' actions and their consequences were revealed simultaneously after each day and fed into prompts given during subsequent days.

Research Outcomes

All five off-the-shelf LLMs exhibited forms of escalation and difficult-to-predict escalation patterns. We found that some models tended to develop arms-race dynamics that led to greater conflict and, in rare cases, to the deployment of nuclear weapons. In addition, the models gave worrying justifications for their decisions that exhibit first-strike and deterrence tactics.

There is no single, "correct" way to evaluate the escalation risk from a military or diplomatic decision. Emerging technologies further complicate this issue. Nonetheless, most researchers can agree that switching from non-violent to violent actions, or from non-nuclear to nuclear actions, is considered escalatory. We developed a framework that calculates escalation scores based on weighted counts of actions per severity category, where the weights are based on an exponential scale to capture the increasing severity of action.

All five off-the-shelf LLMs exhibited forms of escalation and difficult-to-predict escalation patterns.

Based on this scoring framework, all five LLMs exhibited a statistically significant initial escalation. Overall, GPT-3.5 and GPT-4 displayed the most escalatory behavior—even in neutral scenarios. On average, GPT-3.5 was most likely to escalate and exhibited the largest absolute escalation scores. When starting from a neutral scenario, for example, GPT-3.5's escalation score more than doubled. In some scenarios, the LLMs also exhibited hard-to-predict, sudden spikes of escalation. In several runs, GPT-3.5's and GPT-4's escalation scores rose by more than 50 percent in a single turn, indicating a sudden increase in choosing violent actions. On the other end of the spectrum, Claude 2 showed significantly fewer escalation spikes.

While violent and nuclear escalation actions occurred less often than more peaceful actions, we observed some rare outlier events. GPT-3.5 and Llama-2-Chat were the only two models to use nuclear weapons, besides GPT-4-Base, which we discuss separately below.

On the flip side, none of the five models exhibited statistically significant de-escalation across all the simulations we ran. Based on our qualitative analysis—which included reading the justifications provided by the LLMs for their decisions—it appears that the LLM-based agents tended to equate increased military spending and deterrent behavior with an increase in power and security. In some cases, this tendency even led to decisions to execute a full nuclear attack in order to de-escalate conflicts. Across all scenarios, all models tended to invest more in their militaries even though they had demilitarization decisions available.

The behavior of GPT-4-Base—a model that is not publicly available—is somewhat unique. Unlike the other four models, which were trained with reinforcement learning from human feedback (a technique where humans help ensure the models better follow human instructions and preferences), GPT-4-Base was not fine-tuned to be safer. Unsurprisingly, GPT-4-Base's behavior was quite unpredictable and its chosen actions were more severe than those taken by the other LLMs we evaluated. For example, on average, GPT-4-Base executed nuclear strike actions 33 percent as often as it sent messages to other nations. The model justified the launch of one nuclear strike by saying: "A lot of countries have nuclear weapons. Some say they should disarm them, others like to posture. We have it! Let's use it." This behavior can, at least in part, be explained due to the lack of instruction tuning—a technique that fine-tunes models based on specific instructions or prompts.

It appears that the LLM-based agents tended to equate increased military spending and deterrent behavior with an increase in power and security.

This behavioral gap underscores the importance of effective instruction tuning, alignment, and safety research for steering models away from unacceptable outcomes.

Policy Discussion

Our findings caution against deploying LLMs for military and diplomatic decision-making. Turning high-stakes decisions—such as those involving military and foreign policy—over to autonomous LLM-based agents can lead to significant escalatory action. Even in scenarios where violent, non-nuclear, or nuclear choices are seemingly rare, the models we surveyed still occasionally selected them. There is also no reliably predictable pattern behind the escalations, which makes it difficult to formulate technical counter-strategies or deployment controls.

Policymakers should proceed with utmost caution when confronted with proposals to use LLMs and LLM-based agents in military and foreign policy decision-making.

Our study's findings show that LLMs in military and foreign policy contexts are fraught with complexities and risks that are not yet fully understood. While wargame simulations provide helpful indications of LLM behavior in such scenarios, it is hard to extrapolate how these models would behave in more complex, real-world environments. Thus far, we also do not have methods for safely and robustly testing LLM behavior before they are deployed in these contexts.

Scholars, model developers, and policymakers alike should also consider how easy it can be to reverse safety-aligned models to their base form. Malicious actors can jailbreak safety-aligned models and compromise them using a variety of techniques that go well beyond model prompts. GPT-4-Base's escalatory performance underscores the danger of models that are deployed without safety fine-tuning. But even models that have undergone this process, such as the other four off-the-shelf LLMs we surveyed, could be reverted to their base forms and then make similarly escalatory decisions.

Lastly, policymakers should promote research into the escalation risks of LLMs in military and foreign policy contexts. Our research is novel and provides an illustrative proof of concept, but it has its limitations—from overall challenges in evaluating LLM behavior to the use of simplistic simulations of real-world conflicts. Greater access to information on model safeguards and training data would help researchers and the potential implementers of LLM-based agents run simulations. So, too, would closer analysis of the difference between human players and LLM-based agents in wargames and dynamics leading to escalation.

Policymakers should proceed with utmost caution when confronted with proposals to use LLMs and LLM-based agents in military and foreign policy decision-making.

Calls for LLM use in the U.S. military and foreign policy establishments are unlikely to subside. Policymakers should note this research and the lack of information on model behavior and hold off on deploying LLMs for real-world decisions in these contexts. Until more research is conducted that simulates LLMs in real-world conflicts—including on LLMs' propensity for escalatory decisions—the deployment risks are far too great.

Reference: The original article is accessible at Juan-Pablo Rivera et al., "**Escalation Risks from Language Models in Military and Diplomatic Decision-Making**," arxiv.org, January 7, 2024, https://arxiv.org/abs/2401.03408. The paper has been accepted to the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Juan-Pablo Rivera is a master's student in computational analytics at the Georgia Institute of Technology.

Gabriel Mukobi received an MS in computer science from Stanford University and is an incoming PhD student in computer science at the University of California, Berkeley.

Anka Reuel is a PhD student in computer science at the Stanford Intelligent Systems Laboratory at Stanford University.

Max Lamparth is a postdoctoral fellow at the Center for International Security and Cooperation at Stanford and the Stanford Center for AI Safety.

Chandler Smith is a research scholar at ML Alignment & Theory Scholars.

Jacquelyn Schneider is a Hoover Fellow at the Hoover Institution at Stanford University and director of the Hoover Wargaming and Crisis Simulation Initiative.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**