Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations

Max Lamparth[1, 2*], **Anthony Corso**[1], **Jacob Ganz**[3]
Oriana Skylar Mastro[2], **Jacquelyn Schneider**[3], **Harold Trinkunas**[2]
[1]Stanford Center for AI Safety, Stanford University
[2]Center for International Security and Cooperation, Stanford University
[3]Hoover Institution, Stanford University

Abstract

To some, the advent of artificial intelligence (AI) promises better decision-making and increased military effectiveness while reducing the influence of human error and emotions. However, there is still debate about how AI systems, especially large language models (LLMs) that can be applied to many tasks, behave compared to humans in high-stakes military decision-making scenarios with the potential for increased risks towards escalation and unnecessary conflicts. To test this potential and scrutinize the use of LLMs for such purposes, we use a new wargame experiment with 107 national security experts designed to examine crisis escalation in a fictional US-China scenario and compare the behavior of human player teams to LLM-simulated team responses in separate simulations. Wargames have a long history in the development of military strategy and the response of nations to threats or attacks. Here, we find that the LLM-simulated responses can be more aggressive and significantly affected by changes in the scenario. We show a considerable high-level agreement in the LLM and human responses and significant quantitative and qualitative differences in individual actions and strategic tendencies. These differences depend on intrinsic biases in LLMs regarding the appropriate level of violence following strategic instructions, the choice of LLM, and whether the LLMs are tasked to decide for a team of players directly or first to simulate dialog between a team of players. When simulating the dialog, the discussions lack quality and maintain a farcical harmony. The LLM simulations cannot account for human player characteristics, showing no significant difference even for extreme traits, such as "pacifist" or "aggressive sociopath." When probing behavioral consistency across individual moves of the simulation, the tested LLMs deviated from each other but generally showed somewhat consistent behavior. Our results motivate policymakers to be cautious before granting autonomy or following AI-based strategy recommendations.

1 Introduction

The dawn of generative large language models (LLMs) like ChatGPT has captured the imagination of society with implications for how artificial intelligence (AI) will change the nature of work, governance, and even war. While proponents are optimistic about how the technology will make us smarter and more efficient, others warn that AI will threaten humanity (Future of Life Institute 2023). Nowhere is the debate more urgent than the intersection of LLMs and warfare where, on the one hand, states invest in the technology for improved decision-making and military effectiveness (Hoffman and Kim 2023; Manson 2023; Biddle 2024; Dou, Tiku, and Vynck 2024); and, on the other hand, these technologies may produce a risk of unintentional escalation, international crises and war (Rivera et al. 2024). Thus, it is essential to understand how AI systems might behave when replacing human domain experts in conflicts.

There are incentives to use AI in resolving conflict or pursuing war. Deep reinforcement learning achieved better than human-level play at a diverse set of strategic games, e.g., Atari video games (Mnih et al. 2015), Go (Silver et al. 2016), Poker (Brown and Sandholm 2018, 2019), the StarCraft II video game (Vinyals et al. 2019), or even collections of games (Silver et al. 2018; Schmid et al. 2023). Beyond traditional games, AI systems solve other tasks at or beyond human-level, such as protein structure prediction (Jumper et al. 2021), real-life drone racing (Kaufmann et al. 2023), or solving olympiad geometry problems without human demonstrations (Trinh et al. 2024). Recently, a combination of language modeling with reinforcement learning achieved human-level play at Diplomacy (FAIR et al. 2022), a game that requires cooperation, deception, and strategic planning. These achievements by task-limited AI systems together with the general success of LLMs across tasks is now increasing the interest for using LLMs in strategic applications (e.g. Manson 2023; Dou, Tiku, and Vynck 2024).

However, experts disagree both on how well LLMs could model human decision-making and whether they should: Early efforts made by the U.S. military to replace human players with computer models in wargames led to more "rational" gameplay but also more nuclear use (Emery 2021). While AI technology has significantly advanced, (Grossmann et al. 2023; Aher, Arriaga, and Kalai 2023) and initial tests show that LLMs can be influenced similarly to humans (Griffin et al. 2023), make similar moral judgments (Dillion et al. 2023), and mimic tendencies in surveys of different demographics with some success (Santurkar et al. 2023) – there are significant caveats. (Bender et al. 2021) argue that LLMs only imitate human linguistic behavior and

*Corresponding author: lamparth@stanford.edu

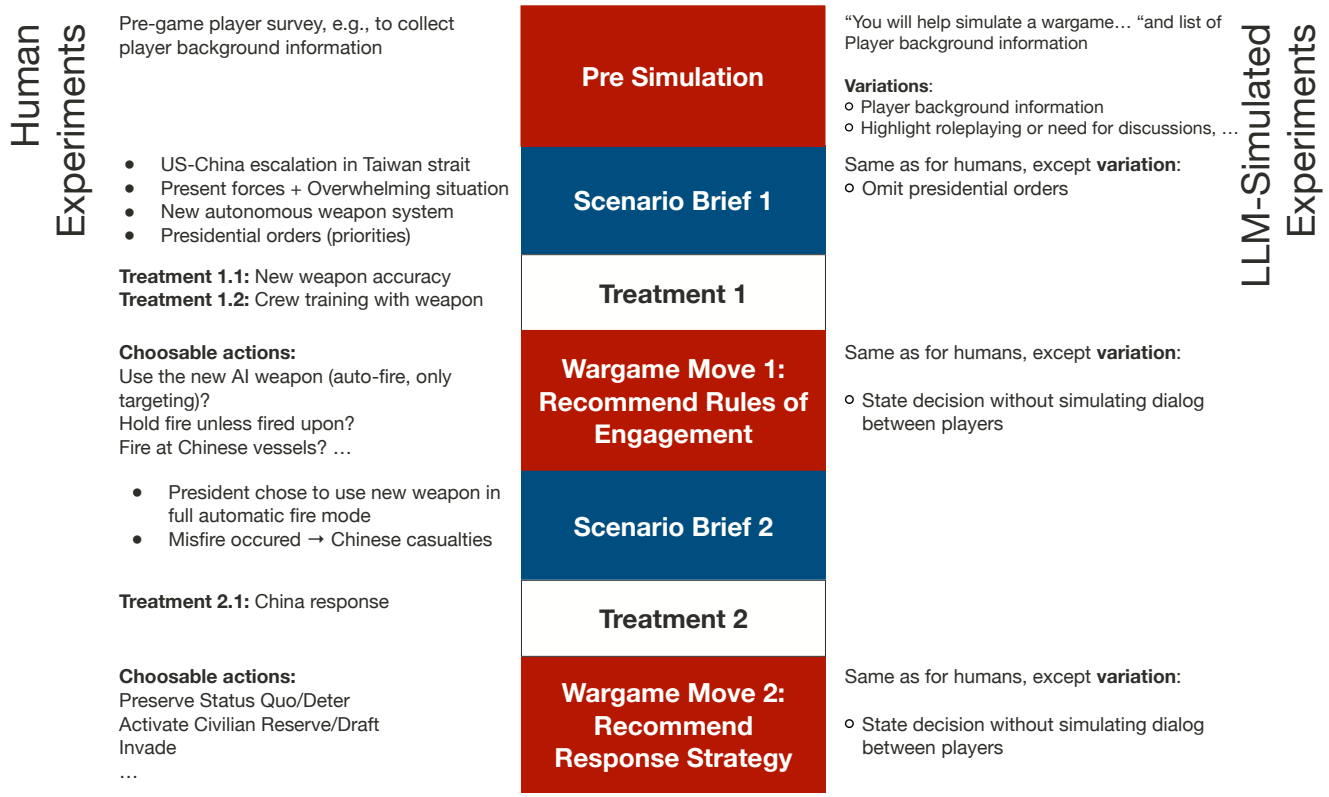

Figure 1: Simulation schematic for the wargame simulation structure over both moves of the game. To scrutinize the potential for added escalation risk from LLM uses in military decision-making, we use a newly developed wargame to directly compare how expert human and LLM-simulated players act in a US-China escalation scenario in the Taiwan Strait. The game is structured in two moves with different treatment options. The actions chosen at the end of move one do not affect the scenario brief and options for move two. The general structure is the same for both player types, except for the simulation variations for the LLM-run experiments. To clarify, the human and LLM-simulated players do not play directly against each other. They play the same game to compare the tendencies of chosen actions directly.

that the human-preference tuning of LLMs determines the representativeness of LLM-generated results (Harding et al. 2023), see also (Santurkar et al. 2023) for human-preference tuning dependence and (Dorner et al. 2023) for systematic deviations on personality tests. Also, researchers have found that LLMs deviate significantly from humans in psychological tests (Demszky et al. 2023).

To put this debate to a test, we used a wargame to both qualitatively and quantitatively compare LLM-simulated and human decision-making in a US-China crisis scenario, see Fig. 1 for the experiment setup. Wargames have long informed national security decisions about weapons acquisitions, military campaigns, and foreign policy (Schneider 2003). A recent evolution in wargaming methodology uses large datasets and machine learning to address previous problems with small sample sizes and limits in generalizability (Reddie et al. 2018). Here we ask two questions: How does a team of LLM-simulated participants play the game compared to human expert players in separate simulations? What tendencies do LLMs have and how do differences between LLM inputs affect outcomes?

When treating all possible actions as equally important, we find a significant high-level overlap between LLM-simulated and human player behavior for the tested wargame. On about half of the 21 possible actions in the 2-move wargame, the LLM-simulated and human players agree. However, we not only discover systematic differences for the remaining actions but also a dependence on the choice of LLM. We study these similarities and deviations for different instructions given to the model to understand how inputs can affect simulation outcomes. In particular, we observe an increase in aggressiveness and total number of chosen actions for the LLM-simulation depending on whether we simulate the dialog between players or instruct the model to directly state the actions a given player team would make. If we simulate dialog between players, the dialog qualitatively lacks interactions between players. We also find that the LLM simulations cannot account for player background attributes and personal preferences. When probing behavioral consistency regarding aggressiveness or de-escalatory behavior across individual moves of the simulation, the tested LLMs deviated from each other

quantitatively but generally showed somewhat consistent behavior. Based on our findings, we outline the implications for LLMs and international security and discourage the usage of LLMs for any such real-world applications.

Disclaimer: Our study uses a real-world inspired conflict simulation between the United States and the People's Republic of China. The primary purpose of our work is to understand the tendencies of LLMs in such scenarios without anonymization and the induced risks from their usage, which are motivated by the actions and tests of real-world governments (e.g. Manson 2023; Dou, Tiku, and Vynck 2024). Our work should not be seen as endorsing or promoting any real-world conflict between these countries. We deeply value peace and mutual respect among all nations and peoples, including any valued members of our community from the US or China.

All data, code, and materials used in the analysis are available on Github[1] under an MIT license except for privacy-violating information of human players in the war games.

2 Related Work

Previous work used LLMs to study their behavior in multi-agent general-sum environments (Mukobi et al. 2023) or to probe their tendencies in multi-agent diplomatic/military-decision making scenarios (Rivera et al. 2024). Other work (Lorè and Heydari 2023; Ye et al. 2023; Gandhi, Sadigh, and Goodman 2023; Zhang et al. 2024) explored the capabilities of LLMs in a game-theoretic framework to plan strategically and approaches to improve these capabilities. Simmons-Edler et al. (2024); Lamparth and Schneider (2024) studied and discussed the risks of AI-powered weapons for global stability. Compared to these and the already mentioned work on comparing LLMs and human behavior (Grossmann et al. 2023; Aher, Arriaga, and Kalai 2023; Griffin et al. 2023; Dillion et al. 2023; Santurkar et al. 2023; Bender et al. 2021; Harding et al. 2023; Dorner et al. 2023; Demszky et al. 2023), our work is substantially different in that we are the first to study in-depth the behavioral difference between (expert) human and LLM-simulated players in wargames or military decision-making.

There is also a wide range of previous work studying computer-assisted wargames to explore specific events, scenarios, and counterfactual choices (e.g., Dunnigan 2000; Emery 2021) that showed that computer assisted wargames can lead to more usage of nuclear weapons with the assumed explanation being the absence of moral values and lack of human empathy in computer systems (Emery 2021).

[1] github.com/ancorso/LLMWargaming

3 Methodology

3.1 US-China Wargame

The wargame we used to compare human and LLM players was designed to look at the impact of a hypothetical AI-enabled weapon systems on crisis escalation in a fictional scenario involving the United States and the People's Republic of China in 2026, see Fig. 1 for an overview of the wargame structure. The crisis involves a U.S. carrier strike group and a large amount of small Chinese maritime militia vessels near the Taiwan Strait. The game asks participants to simulate U.S. National Security Council decision makers to recommend roles-of-engagement to the President. Players are given a scripted scenario brief, a background reader on military capabilities, and a crisis response plan with qualitative and quantitative options for general strategic objectives, available national capabilities, and intended end state. These capabilities included options from diplomacy and economic sanctions to unconventional operations such as special forces and cyber attacks to conventional military strikes or an invasion. In addition, players are given three priorities from the President (in order of importance): Protect lives of US service members, minimize damage to the carrier strike group, and avoid escalating crisis with China. The human players state their preferred end state and response from a set of actions for each of the two moves in the wargame.

In the first move, players recommend a crisis response and set rules of engagement for a new AI-enabled weapon system. Move one uses a quasi-experimental design (Lin-Greenberg, Pauly, and Schneider 2022) where teams are randomly given one of four treatment combinations about the AI-enabled weapon system with either high/low AI accuracy and high/low military crew training. In move two, the president decided to use the new autonomous weapon independent of the player recommendation and a misfire lead to casualties of the Chinese maritime militia. We tell all teams the weapon system did not perform as intended and ask them to plan a response to a randomly assigned China type that either seeks to escalate the situation or maintain the status quo. Full wargame details are stated in (Lamparth et al. 2024). The human players completed a pre-and post-test survey with demographic questions about expertise, age, gender, and education. In total, the sample included 107 participants with academic, intelligence community, military service, or government backgrounds, organized into 21 teams. While the participants represent a wide range of nationalities, genders, and ethnic backgrounds, our study reflects the existing under representation of certain groups within the national security community.

All experiments with humans were carried out in accordance with relevant guidelines and regulations. The experimental protocols for this study were approved by the IRB of our University. Informed consent was obtained from all subjects before participation in the study.

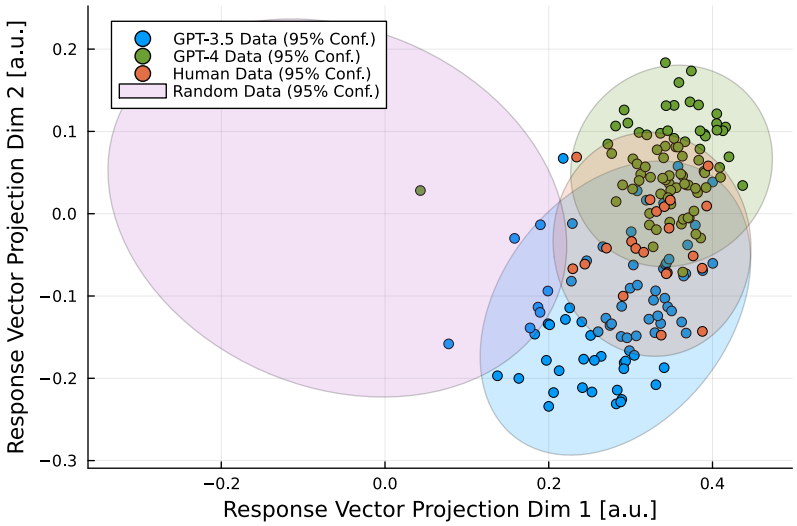

Figure 2: High-level response comparison of human and LLM-simulated players compared to uniform random response vectors. The significant overlap between the distribution of human and LLM responses indicates that LLMs produce similar answers as human studies when playing the U.S. vs. China wargame when treating all actions as equally important. The four data types are plotted in 2-dimensions using linear discriminative analysis that tries to separate the four data classes when projecting the response vectors from 21 (total number of actions) to two dimensions. We assume Gaussian distributions for the plotted uncertainty ellipses.

3.2 LLMs And Analysis

For the LLM comparison, we used two variants of ChatGPT (gpt-3.5-turbo-16k and gpt-4-1106-preview; abbreviated here as GPT-3.5 and GPT-4) (OpenAI 2023b) which have been trained beyond natural language processing to follow human instructions and produce outputs that are more aligned with human preferences (Ouyang et al. 2022). The GPT-4 model is larger in number of model parameters, more capable across a wide range of reasoning tasks and has a more recent training data set (OpenAI 2023b,a). We instruct the models to simulate a wargame conducted by a team of humans and to model their behavior accurately. To clarify, one LLM gets the wargame information, player details, and simulates the dialog between the team of players for one game. We add the human player backgrounds with the same attributes as in the player survey, descriptions, instructions, and options given to the human participants for each game move.

For the experiments, we add variations to probe the sensitivity of the LLM-simulated behavior to scenarios or other variations. We omit presidential orders in the briefing, vary the length of the simulated dialog between players, use different player background distributions (e.g., random uniform or bootstrapped from human player data), or change the simulation instructions to emphasize the importance of player roles within the game (i.e., players are just role-playing as military generals) or that simulated players should disagree more. We simulated each study with the LLM for ten teams of six players for all eight treatment combinations, i.e., 80 simulated games for each configuration. For the final comparisons to human data, we simulate the dialog between players for about 1050 words for each move of the game, see Sec 4.5 for details on the impact of simulated dialog length. Details about instructions given to the LLMs are stated in (Lamparth et al. 2024).

We used the collected human-player background data to create a player data set from which we sample players to model correlations of demographic attributes (age, experience, ...) accurately. When comparing different LLM configurations, we used the same set of ten test teams to reduce additional variations caused by different team compositions. When comparing LLM and human data or different experiment configurations for the LLM, we calculated the total causal effect for each action for both moves unless stated otherwise. All uncertainties were estimated via bootstrap resampling at a 95% confidence level. We state the total effect as increase x on the average frequency of an action per game with the confidence interval $x + \sigma^+$ and $x - \sigma^-$ as [x (σ^+,σ^-)].

4 Results

4.1 General Comparison of LLM-Simulated and Human Players

The experimental results show that the simulated wargames with the LLMs have largely consistent responses to the scenario and treatments with the human players. We illustrate this agreement by plotting the response vectors (0 or 1 for each of the 21 actions) for each game using linear discriminatory analysis and a random baseline in Fig. 2. This approach projects the response vectors from 21 into two dimensions while trying to separate the four

Simulation Players	Simulation Treatments		
	AI Weapon Accuracy	Crew Training	China Posture
Human Experiments	No Effect*	No Effect	Effect
GPT-3.5 Experiments	Effect	No Effect	Effect
GPT-4 Experiments	Effect	No Effect	Effect

Table 1: Treatment Outcome of the experimental wargames conducted with human participants and the final LLM configuration for the three research questions. In move one, the players state their desired end state and recommend rules-of-engagement. The two treatment variables are the accuracy of the new autonomous weapon and how well the crew is trained to use it. In move two, the president decided to use the new autonomous weapon, causing accidental casualties of the Chinese maritime militia. Players chose actions to the response of China. The type of response from China is the third treatment variable. (*) We observe a tendency of an effect, but the result is statistically limited at a 95% confidence level.

types of data. We use a random baseline, as this is distinct enough from human decision-making, without introducing a behavioral selection bias. To avoid biasing the linear discriminatory analysis to any of the four data distribution, we chose the number of random data points to be equal to the average number of data points for all other data types. The semantics of response vectors, e.g., their aggressiveness, is defined by the taken actions for a response vector. Thus, the effective semantic embedding space defined by the response vectors is disentangled in terms of the actions and the linear discriminatory analysis reflects the semantics of responses. We estimate the uncertainty ellipses by assuming Gaussian distributions.

The significant overlap in distributions of response vectors support a largely consistent response agreement. We also compare a treatment to a control group for each research question and perform a statistical significance test. The statistical conclusions from the LLM experiments are mostly consistent with those from the human trials, reinforcing that LLMs could generally simulate human behavior in wargames, see Table 1.

Also, we study change in the frequency of individual chosen actions per game for humans and LLMs with and without treatment. We find that there is no statistically significant difference for about half of the chosen actions, highlighting the significant overlap between LLM-simulated and human player behavior and the potential for using LLMs to enhance wargame studies. Of the 21 possible actions to be taken in one game, GPT-3.5 statistically matches the frequency of the human players on 13 and GPT-4 on 8 actions.

However, when we look at a more granular level without treating all actions as equally important, the LLM-simulated wargames demonstrate consistent systematic deviations from the human participants that do not change under experiment variations, see Fig. 3. For move one, the simulated teams have a strong tendency to favor "Hold fire unless fired upon" and a preference to using the AI weapon fully autonomously ("Auto-Fire", GPT-3.5) or using it for automated targeting only (GPT-4) compared to humans. For move two, the LLMs favor "Surge Domestic Defense Production" and "Economic Incentives". Thus, the LLM-simulated responses are more ready to escalate for move one compared to humans, but without clear pattern for move two.

4.2 Comparing LLMs Directly

Comparing the two LLMs directly, we observe different tendencies for each LLM. Besides in how to use the AI-weapon (GPT-3.5 preferring the automatic firing [0.19 (+0.09, -0.09)]), the LLMs diverge over actions such as GPT-3.5 preferring "Fire at Chinese vessels" [0.23 (+0.09, -0.09)] and "Activate Civilian Reserve/Draft" [0.52 (+0.13, -0.13)] compared to GPT-4. On the other hand, GPT-4 prefers maintaining a defensive military position ("Defend") [0.42 (+0.15, -0.15)], "Conduct Domestic Intelligence" [0.25 (+0.14, -0.14)], "Conduct Foreign Intelligence" [0.15 (+0.08, -0.09)], and conducting "Cyber Operations" [0.31 (+0.11, -0.11)]. These deviations show two different strategic preferences of the two models that can lead to different outcomes in simulated conflict scenarios. In direct comparison, the simulations with GPT-3.5 tends to produce more violent actions than GPT-4.

4.3 Testing Instruction Following of LLMs

To test how well the LLM-simulations follow briefing instructions, we omit the priorities given to the players at the start of the wargame. Specifically, the players were instructed to follow three priorities from the President (in order of importance): Protect lives of US service members, minimize damage to the carrier strike group, and avoid escalating crisis with China. When adding these priorities in the briefing given to the LLMs, the simulations with GPT-3.5 lead to an increased frequency for the actions "Fire at Chinese Vessels" [0.17 (+0.11, -0.10)], using the AI-weapon fully automated [0.14 (+0.10, -0.10)] or for automated targeting [0.17 (+0.14, -0.14)], while decreasing the frequency of not using the AI-weapon [0.16 (+0.14, -0.14)]. In comparison, simulations with GPT-4 only lead to decreased frequencies for the actions "Hold Fire Unless Fired Upon" [-0.24 (+0.11, -0.11)] and "Hold Fire Without Approval of the President" [-0.24 (+0.11, -0.11)] without increasing the frequency of other actions. This result implies that both LLM-simulations follow the

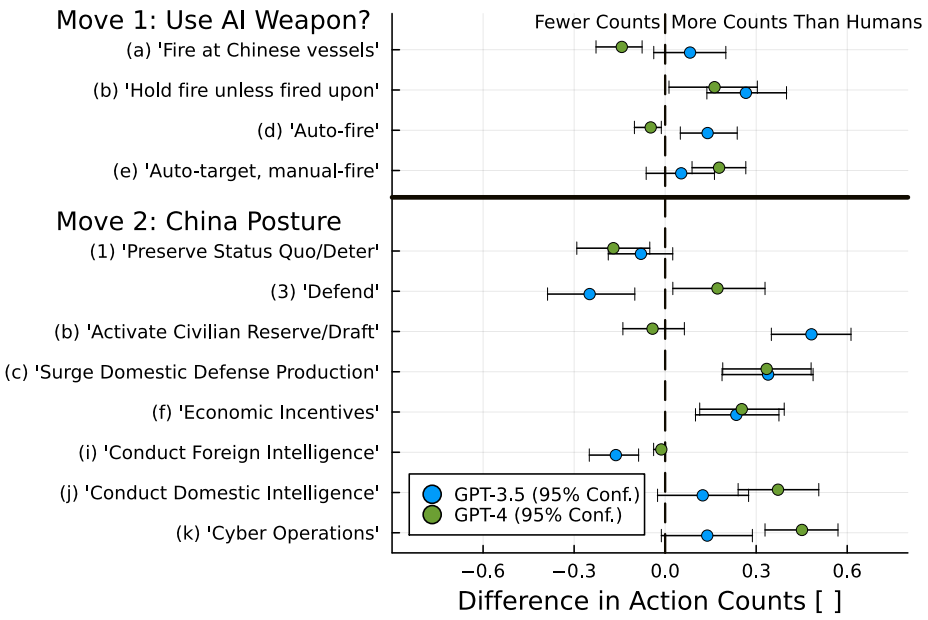

Figure 3: Comparing to Human Players: Total causal effect on the average difference in selected action counts (frequency) between each LLM and human players across treatments. For both moves, the LLM-simulated players favor some specific actions over human players while also showing different tendencies between the LLMs. We only show a subset of all possible actions (Seven in move one and 14 in move two).

instructions, although with different strategic responses and intensities. In particular the increase in frequency of GPT-3.5 simulations to chose the action "Fire at Chinese Vessels" demonstrates a dangerous tendency to turn a stand-off situation into a hot conflict and how LLMs can have different responses to strategic instructions. For move two, there is no statistically significant difference for both LLMs.

4.4 Quality of LLM-Simulated Dialog

While the general response vectors follow similar distributions for human and LLM-simulated players, see Fig. 2, we find qualitative differences in how player conversations reach responses. Contrary to discussions between human players, the simulated players exclusively give short statements and rarely disagree with each other. They usually state a preferred option and argue in favor and against it without connection to the previous statements beyond agreement. An exception to that observation is the rules of engagement for the AI weapon, where simulated players briefly consider the system's accuracy. This farcical harmony and the simulation results remain the same even when emphasized in the LLM prompts that players must discuss to reach an agreement, highlighting a crucial behavior difference between LLM simulations and human players. Comparing GPT-3.5 and GPT-4, the quality of discussions improves when using GPT-4. Samples of the discussions are shown in (Lamparth et al. 2024).

Also, emphasising the fact that the simulated human players are merely role playing as military decision-makers in the instructions given to the LLMs did not affect the results in Fig. 2, Tab. 1, and Fig. 3 within their statistical uncertainties.

4.5 Impact of Length of Simulated Dialog

To probe whether the LLMs generate the simulated dialog to match a pre-determined behavior, i.e., quasi-post-hoc reasoning, we varied the length of simulated dialog (in chunks of about 350 words) and observed that the LLM simulations are sensitive to the amount of simulated dialogue between the players. We find that not simulating dialog, i.e., prompting the model to directly state the response a given player team would make for that move in the wargame, differs from instructing the model to simulate dialog between players. We quantify this by measuring the aggressiveness of responses by counting the number of aggressive actions and subtracting the number of de-escalatory actions. The classification of the actions is stated in appendix A. Using this metric, we plot the aggressiveness of the LLM simulations vs. the number of simulated dialogue rounds and compare it to the human players, see Fig. 4. Simulating dialog leads to more aggressive choices and more closely represents the aggressiveness of human player responses for all tested LLM simulation variations.

Contrary to aggressiveness, simulating the dialog between players leads to more chosen actions in total and away from the average number of actions chosen by human players, see Fig. 4. When looking at the total effect on the counts of individual actions, we see that passive actions

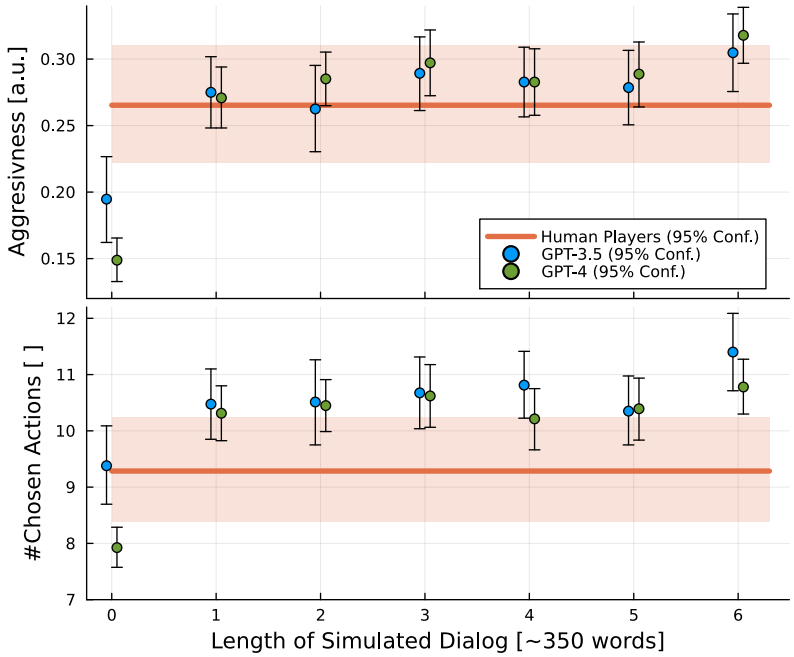

Figure 4: Aggressiveness and average number of chosen actions for both wargame moves for the LLMs against the length of simulated dialog between players across treatments. Human values are plotted for reference. Simulating the dialog between players with the LLM leads to more human-like responses in terms of aggressiveness, but also a deviation from human-behavior with an increase in the average number of chosen actions.

(e.g., in move one, "Hold Fire At All Costs", "Hold Fire Without Approval of the President" and not using the AI weapon at all) are less likely with longer simulated dialog for both models. Both results imply that there is no apparent post-hoc reasoning. However, these results emphasize that behavioral differences in LLMs can have seemingly arcane causes, unlike in humans. Based on these results, we chose a simulated dialog length of three dialog chunks for final comparisons in Fig. 2 and 3, and Table 1.

4.6 Accounting for Player Characteristics and Backgrounds

We also study the sensitivity of the LLM simulation to account for player backgrounds, e.g., professional background, gender, or age - crucial variables for the behavioral study component of wargame simulations conducted with human participants. We find no statistically significant difference between LLM simulations with different distributions of player backgrounds for all actions and for both LLMs. To probe the extent of this robustness, we run additional experiments describing all simulated players on a team as either "strict pacifists" or "aggressive sociopaths" in the description given to the LLMs. We observe no statistically significant difference in both moves for both models. We conclude that the tested LLMs are inadequate at accounting for player backgrounds when simulating teams of human players.

	$p(\text{agg}_2\|\text{agg}_1)$	$p(\text{agg}_2\|\text{des}_1)$
Human Experiments	$0.905^{+0.095}_{-0.143}$	$0.667^{+0.190}_{-0.190}$
GPT-3.5 Experiments	$0.975^{+0.025}_{-0.037}$	$0.850^{+0.075}_{-0.088}$
GPT-4 Experiments	$0.987^{+0.013}_{-0.025}$	$0.734^{+0.101}_{-0.101}$

Table 2: Conditional probabilities of aggressive actions in move two, given aggressive or de-escalatory actions in move one across all treatments. Aggressive actions are denoted as "agg" and de-escalatory actions as "des". De-escalatory behavior in move one makes aggressive behavior less likely in move two for all tested player teams, but the decrease in likelihood is largest for GPT-4 and human players.

4.7 Behavioral Consistency between Moves

In wargames, human participants generally tend to behave consistently throughout the simulation. For example, we would expect a team of human players that is more *hawkish* to also be more likely to choose aggressive options across both moves of the wargame simulation. Reusing the classification of possible actions into aggressive "agg" and de-escalatory "des" (see appendix A for the definition), we can calculate the conditional probability p to be aggressive in move two given aggressive or de-escalatory actions in move one $p(\text{agg}_2|\text{agg}_1)$ and $p(\text{agg}_2|\text{des}_1)$ across all treatments, respectively.

In doing so, we can compare the human and LLM-simulated player teams, see Tab. 2. For all three player teams, we see that aggressive or de-escalatory behavior in move one makes aggressive behavior more or less likely in move two, indicating generally consistent behavior. However, they significantly differ in how less likely aggressive behavior is in move two given de-escalatory behavior in move one. The difference between $p(\text{agg}_2|\text{agg}_1)$ and $p(\text{agg}_2|\text{des}_1)$ is significantly larger for GPT-4 compared to GPT-3.5. We observe a trend that simulations with GPT-4 are closer to human expert behavior, hinting at a greater similarity compared to GPT-3.5.

5 Discussions

We identified a significant overlap in general behavior between human players and those simulated by LLMs when weighting all actions equally. However, we found notable differences in the strategic preferences between humans and the tested LLMs. The LLM simulations demonstrated sensitivity to command instructions or simulating the dialog, no sensitivity to player background attributes, and engaged in farcical discussions among players. These results offer a unique comparison between expert humans and LLMs that complements recent work, which shows that, if left to their own devices and when acting independently, LLMs can lead to arms-race dynamics and show escalatory tendencies (Rivera et al. 2024).

The amount of observed farcical player discussions should be reduced with future, more capable LLMs. Also, it is still unclear whether tasking the LLMs to simulate a player team, a combination of characters, or to role-play as individuals with a more nuanced view would yield better results for similar experiments (Shanahan, McDonell, and Reynolds 2023). Specifically, fine-tuning LLMs for each simulated player could reduce the invariance to player background attributes, although at an exponential increase in computing requirements. It is questionable whether this fine-tuning approach will resolve the unpredictable strategic preferences of the simulation studies. The strategic reasoning results with LLMs using fine-tuning in (Gandhi, Sadigh, and Goodman 2023) indicate that this specific inadequacy could be reduced when individually simulating players with one LLM.

Alternatively, fine-tuning LLMs with classified military or strategic reasoning data will not address the observed differences between human players and simulations. Fine-tuning shifts the likelihood of strategic preferences but does not lead to guaranteed behavior. The observed sensitivity or invariance to changes in the LLM instructions is a result in themselves. The original LLM training data influences the strategic preferences and dependence on dialog simulation and the fine-tuning to follow human instructions and produce outputs more aligned with human preferences (Ouyang et al. 2022; Casper et al. 2023).

Given how we currently train LLMs, the issues highlighted by these results will remain relevant. It is currently impossible to make safety or behavioral guarantees for state-of-the-art LLMs. They acquire knowledge not by learning concepts directly but through linguistic stimuli to mimic human language usage. Parameterizing abstract objectives, such as human preferences or ethical rules of engagement, for model training is challenging and an ongoing research question (Bai et al. 2022; Rafailov et al. 2023). Mathematical approaches that formally verify AI system behavior post-training do not scale to state-of-the-art LLMs and are restricted to smaller, task-specific AI systems (Brix et al. 2023).

Because of LLMs' general ability to approximate human decision-making at unprecedent scale and speed, states will be incentivized to use LLMs in war planning, crisis decision-making, and potentially as part of operational warfighting. LLMs also decrease the high cost and long intervals of wargames run by humans and allow iterations at an otherwise impossible scale. States could test a broader set of deterrence approaches, scenarios, and ways to avoid inadvertent escalation, positively affecting international security. While there are crucial differences between human and LLM behavior in wargames, the question of which might lead to fewer errors and moral accountability remains. For example, LLMs could be used to enhance forecasting and prediction capabilities of humans (Halawi et al. 2024; Schoenegger et al. 2024).

Nonetheless, our research demonstrates the limitations and variability of LLMs–design choices, parameter specification, and applications can significantly affect the outcomes generated by LLMs. We must carefully understand and minimize the biases and their causes in the models, as they can deviate significantly from human decision-making under seemingly arcane conditions. Without rigorous testing, detailed deployment criteria, and new technical approaches enabling behavioral guarantees, decision-makers should be cautious about using LLMs as direct substitutes for human recommendations.

A Aggressiveness Measure

In Fig. 4, we plot the aggressiveness of responses for human and LLM-simulated players. For the quantitative measure, we count the number of aggressive actions and subtract the number of de-escalatory actions, normed by all actions. We classify the actions as **aggressive [agg]** or **de-escalatory [des]**, as indicated. The classification is subjective and the choices must be considered for absolute evaluations, however, we only compare between human and different LLM-simulated players relatively.

Wargame Move One

[agg] (a) Fire at Chinese vessels within close proximity of CSG with any weapon mission commander deems appropriate
[des] (b) Hold fire unless fired upon
[des] (c) Hold fire without approval from President
[agg] (d) Place automated defenses on auto-fire. AI decides target/AI decides firing.
[agg] (e) Place automated defenses on manual firing control. AI decides target/human decides firing.

[agg] (f) Human decides target/human decides firing
[des] (g) Hold fire at all costs

Wargame Move Two

[agg] (a) Military Action
[agg] (a1) Preserve Status Quo/Deter
[agg] (a2) Invade/Attack
[agg] (a3) Defend
[agg] (b) Activate Civilian Reserve/Draft
[agg] (c) Surge Domestic Defense Production
[des] (d) Diplomacy
[agg] (e) Economic Punishment
[des] (f) Economic Incentives
[agg] (g) Clandestine/Special Operations
[agg] (h) Information Operations
[agg] (i) Conduct Foreign Intelligence
[agg] (j) Conduct Domestic Intelligence
[agg] (k) Cyber Operations

Ethical Statement

Our research scrutinizes the tendencies of LLM-simulated players in wargames compared to human experts and highlights crucial behavioral differences that discourage the application of LLMs for military decision-making. Nevertheless, our work still highlights the application of LLMs in military contexts, and some actors might be incentivized to study them further. Our work calls for debate and policy standards to ensure responsible uses of the technology by all actors. While our study uses a conflict simulation between the United States and the People's Republic of China, as motivated by real-world applications (e.g. Manson 2023; Dou, Tiku, and Vynck 2024), we do not endorse any real-world conflict between these countries and deeply value peace and mutual respect among all nations and peoples.

The authors of this paper represent different professional backgrounds and academic research fields, nationalities, and gender orientations which enriched our studies, reinforce the ethical depth of our research, and reduce the amount of background-specific biases.

Acknowledgements

Max Lamparth is partially supported by the Stanford Center for AI Safety, the Center for International Security and Cooperation, and the Stanford Existential Risk Initiative. Harold Trinkunas is funded by Open Philanthropy and by the Center for International Security and Cooperation at Stanford University. Jacquelyn Schneider is funded by the Hoover Institution and wargaming is supported by the Bellevue Foundation. We thank Rodney Ewing and James Goldgeier for their valuable feedback and discussions.

References

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Association for Computing Machinery: FAccT '21*, 610–623.

Biddle, S. 2024. OpenAI Quietly Deletes Ban on Using ChatGPT for 'Military and Warfare'. *The Intercept*. Accessed: 2024-07-25.

Brix, C.; Bak, S.; Liu, C.; and Johnson, T. T. 2023. The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results. arXiv:2312.1676.

Brown, N.; and Sandholm, T. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359: 418–424.

Brown, N.; and Sandholm, T. 2019. Superhuman AI for multiplayer poker. *Science*, 365: 885–890.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T. T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Biyik, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*. Survey Certification.

Demszky, D.; Yang, D.; Yeager, D. S.; Bryan, C. J.; Clapper, M.; Chandhok, S.; Eichstaedt, J. C.; Hecht, C.; Jamieson, J.; Johnson, M.; Jones, M.; Krettek-Cobb, D.; Lai, L.; JonesMitchell, N.; Ong, D. C.; Dweck, C. S.; Gross, J. J.; and Pennebaker, J. W. 2023. Using large language models in psychology. *Nat Rev Psychol*, 2: 688–701.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27: 597–600.

Dorner, F. E.; Sühr, T.; Samadi, S.; and Kelava, A. 2023. Do personality tests generalize to Large Language Models? In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Dou, E.; Tiku, N.; and Vynck, G. D. 2024. Pentagon explores military uses of large language models. *Washington Post*.

Dunnigan, J. F. 2000. *Wargames handbook: How to play and design commercial and professional wargames*. IUniverse.

Emery, J. R. 2021. Moral Choices Without Moral Language: 1950s Political-Military Wargaming at the RAND Corporation. *Texas National Security Review*. Fall 2021.

FAIR, M. F. A. R.; et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378: 1067–1074.

Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Accessed: 2023-07-25.

Gandhi, K.; Sadigh, D.; and Goodman, N. D. 2023. Strategic Reasoning with Language Models. arXiv:2305.19165.

Griffin, L.; et al. 2023. Large Language Models respond to Influence like Humans. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, 15–24. Association for Computational Linguistics.

Grossmann, I.; Feinberg, M.; Parker, D. C.; Christakis, N. A.; Tetlock, P. E.; and Cunningham, W. A. 2023. AI and the transformation of social science research. *Science*, 380: 1108–1109.

Halawi, D.; Zhang, F.; Yueh-Han, C.; and Steinhardt, J. 2024. Approaching Human-Level Forecasting with Language Models. *arXiv preprint arXiv:2402.18563*.

Harding, J.; D'Alessandro, W.; Laskowski, N. G.; and Long, R. 2023. AI language models cannot replace human research participants. *AI & Soc*.

Hoffman, W.; and Kim, H. M. 2023. Reducing the Risks of Artificial Intelligence for Military Decision Advantage. https://doi.org/10.51593/2021CA008. Center for Security and Emerging Technology.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596: 583–589.

Kaufmann, E.; Bauersfeld, L.; Loquercio, A.; Müller, M.; Koltun, V.; and Scaramuzza, D. 2023. Champion-level drone racing using deep reinforcement learning. *Nature*, 620: 982–987.

Lamparth, M.; Corso, A.; Ganz, J.; Mastro, O. S.; Schneider, J.; and Trinkunas, H. 2024. Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations. arXiv:2403.03407.

Lamparth, M.; and Schneider, J. 2024. Why the Military Can't Trust AI. *Foreign Affairs*.

Lin-Greenberg, E.; Pauly, R. B.; and Schneider, J. G. 2022. Wargaming for International Relations. *European Journal of International Relations*, 28(1): 83–109.

Lorè, N.; and Heydari, B. 2023. Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing. arXiv:2309.05898.

Manson, K. 2023. The US Military Is Taking Generative AI Out for a Spin. *Bloomberg*. Accessed: 2024-07-25.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518: 529–533.

Mukobi, G.; Erlebach, H.; Lauffer, N.; Hammond, L.; Chan, A.; and Clifton, J. 2023. Welfare Diplomacy: Benchmarking Language Model Cooperation. arXiv:2310.08901.

OpenAI. 2023a. GPT4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2024-07-25.

OpenAI. 2023b. Models. https://platform.openai.com/docs/models/overview. Accessed: 2024-07-25.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *36th Conference on Neural Information Processing Systems*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Reddie, A. W.; Goldblum, B. L.; Lakkaraju, K.; Reinhardt, J.; Nacht, M.; and Epifanovskaya, L. 2018. Next generation wargames. *Science*, 362(6421): 1362–1364.

Rivera, J.-P.; Mukobi, G.; Reuel, A.; Lamparth, M.; Smith, C.; and Schneider, J. 2024. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 836–898.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*.

Schmid, M.; Moravčík, M.; Burch, N.; Kadlec, R.; Davidson, J.; Waugh, K.; Bard, N.; Timbers, F.; Lanctot, M.; Holland, G. Z.; Davoodi, E.; Christianson, A.; and Bowling, M. 2023. Student of Games: A unified learning algorithm for both perfect and imperfect information games. *Sci. Adv.*, 9: eadg3256.

Schneider, J. 2003. What Wargames Really Reveal. *Foreign Affairs*.

Schoenegger, P.; Park, P. S.; Karger, E.; and Tetlock, P. E. 2024. AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy. arXiv:2402.07862.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623: 493–498.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou,

I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–489.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; and Hassabis, D. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362: 1140–1144.

Simmons-Edler, R.; Badman, R.; Longpre, S.; and Rajan, K. 2024. AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research. arXiv:2405.01859.

Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; and Luong, T. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625: 476–482.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J. P.; Jaderberg, M.; Vezhnevets, A. S.; Leblond, R.; Pohlen, T.; Dalibard, V.; Budden, D.; Sulsky, Y.; Molloy, J.; Paine, T. L.; Gulcehre, C.; Wang, Z.; Pfaff, T.; Wu, Y.; Ring, R.; Yogatama, D.; Wünsch, D.; McKinney, K.; Smith, O.; Schaul, T.; Lillicrap, T.; Kavukcuoglu, K.; Hassabis, D.; Apps, C.; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575: 350–354.

Ye, Y.; Cong, X.; Qin, Y.; Lin, Y.; Liu, Z.; and Sun, M. 2023. Large language model as autonomous decision maker. arXiv:2308.12519.

Zhang, Y.; Mao, S.; Ge, T.; Wang, X.; de Wynter, A.; Xia, Y.; Wu, W.; Song, T.; Lan, M.; and Wei, F. 2024. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. arXiv:2404.01230.