Hawkish Bias in Foreign Policy Decision-Making: Comparative Experiments with Human and LLM Agents *

Yifei Zhu[1], Jingtian Hu[2], Zhenhua Tu[3], Yinzhi Lin[4], and Xun Pang[3,*]

[1]Department of Politics and Public Adniministration, The University of Hong Kong
[2]Department of International Relations, School of Social Science, Tsinghua University
[3]School of International Studies, Peking University
[4]School of Government, Peking University

November 22, 2025

Abstract

Foreign policy decision-making involves unstructured problems, ambiguous goals, and information that is scarce, unreliable, and time-sensitive—conditions under which human decision-makers often rely on cognitive shortcuts and, in turn, display well-documented "hawkish biases." This paper asks whether large language model (LLM)-based agents, when placed in similar decision environments, reproduce or mitigate these tendencies. Replicating three experiments from Kertzer et al. (2022), we construct a multi-agent environment of 3,973 LLM-based agents, each designed to embody a specific bias underlying hawkish foreign policy preferences. Our results show that AI agents are as susceptible to bias as humans, and that multi-agent interaction does not attenuate these tendencies. Yet the character of machine bias differs qualitatively: it is inconsistent, unstable, and unpredictable, likely due to "persona drift," whereby LLM agents–unlike humans–lack stable internal dispositions, causing experimental interventions to interact with and distort their assigned personas. These results suggest that state-of-the-art LLMs are currently ill-suited to serve as synthetic participants in RCTs, as persona instability introduces substantial post-treatment confounding.

Keywords: Cognitive Biases, Foreign Policy Decision-Making, Large Language Models, Multi-Agent System, Causal Inference, Randomized Controlled Experiments

*Preliminary draft. The author marked with an asterisk (*) is the corresponding author. Please direct any comments or questions to Xun Pang at xpang@pku.edu.cn.

1 Introduction

The rapid integration of artificial intelligence (AI) into public life is increasingly blurring the boundary between human judgment and machine reasoning, even in domains long considered the exclusive province of human decision-makers. Recent developments underscore this shift: Albania has experimented with appointing an AI bot to assist a government minister in anti-corruption efforts, and the Swedish Prime Minister has publicly acknowledged using ChatGPT in the course of official duties.[1] These examples signal that AI is no longer peripheral to political practice but is beginning to reshape notions of political authority, expertise, and accountability. As AI systems increasingly participate, directly or indirectly, in political decision processes, a systematic assessment of what kinds of "decision-makers" large language models (LLMs) effectively are has become both timely and essential.

Foreign policy decisions constitute a distinct class of decision-making, differing fundamentally from routine choices. They are marked by unstructured and ambiguous problems, information that is scarce, unreliable, and time-sensitive, and the intense cognitive and emotional pressures characteristic of high-stakes contexts. While it is often assumed that routine decisions rely primarily on heuristics and habit, whereas high-stakes decisions invoke more analytical and rational reasoning , the reality in foreign policy is quite the opposite: the very features of these situations compel decision-makers to depend on cognitive shortcuts, leading to systematic deviations from rational choice .

The development of LLMs raises the possibility of mitigating such biases in such decision environments. LLMs lack subjective cognition and emotional pressures, and may therefore act more "rationally" and consistently. LLMs also possess extensive internal knowledge, can rapidly process complex and unstructured information, perform multi-step reasoning, and adapt outputs to specific contexts. Furthermore, multi-agent systems (MASs) built on LLMs might avoid group pathologies, since LLMs are not subject to social pressure or

[1]The Guardian, Aug. 5, 2025. https://www.theguardian.com/technology/2025/aug/05/chat-gpt-swedish-pm-ulf-kristersson-under-fire-for-using-ai-in-role

conformity dynamics . Yet existing research shows that LLMs themselves display deviation, social-category, and demographic biases, and can reproduce or even amplify human cognitive biases through their training data and reinforcement learning from human feedback (Chen, Hu and Lu, 2025; Malberg et al., 2024; Suri et al., 2023). Emerging evidence further suggests that MASs can develop human-like linguistic norms and collective biases (Ashery, Aiello and Baronchelli, 2025) and remain sensitive to communication structures and institutional configurations (Jin et al., 2025; Zhuge et al., 2023).

Consequently, whether AI decision-makers are more or less prone to "hawkish bias" in foreign policy is ultimately an empirical question that remains unresolved.

This study builds on the recent "behavioral turn" in International Relations (IR), characterized by the rapid expansion of randomized controlled trials (RCTs) examining human cognitive and affective tendencies in political decision-making . RCTs provide an ideal benchmark for detecting, comparing, and interpreting potential machine biases. Because RCTs are inherently simulated environments in which decision scenarios are carefully constructed and participants "role-play: as foreign policy decision-makers, they offer a natural platform for parallel comparisons between human and AI-based decisions. Such comparisons also constitute a rigorous test of whether LLM-driven agents can serve as valid synthetic subjects in RCTs, which is an aspiration widely seen as a promising way to mitigate the well-known limitations of RCTs in political science and other social sciences.

More specifically, we systematically replicate Kertzer et al. (2022) using LLM-based agents in both individual and multi-agent configurations. Kertzer et al. (2022) conducted three experiments examining key cognitive biases in foreign policy decision-making–namely, loss-framing bias, intentionality bias, and reactive devaluation bias. These biases collectively contribute to hawkish tendencies in foreign policy, making compromise more difficult, encouraging risk-seeking behavior, and increasing the likelihood of conflict escalation.

Beyond their theoretical significance, these experiments offer valuable variation for assessing AI biases and their comparability to human behavior. The three studies differ in design,

ranging from high-urgency to low-politics contexts, from highly scarce to relatively abundant information environments, and from well-structured to highly unstructured decision problems. Moreover, by incorporating both individual and group decision-making conditions, the original research provides a coherent framework for comparing decision dynamics in single-agent versus multi-agent systems (MAS).

Strictly adhering to the original human-subject experimental designs, we construct a multi-agent environment and assign personas to 3,973 LLM-based agents based on a pre-treatment survey from Kertzer et al. (2022). These agents are randomly assigned to treatment and control groups across three decision-making settings: individual decisions, horizontal group decisions (with equal participation), and hierarchical group decisions (with one leader and several advisors). We then compare the average treatment effects (ATEs) from these replications with those observed in human experiments.

The results reveal a complex pattern of machine biases. AI agents reproduce the loss-framing bias, reverse the intentionality bias, and exhibit the reactive devaluation bias. Multi-agent systems (MASs) neither mitigate these biases nor generate average treatment effects (ATEs) that more closely align with human results, nor do they display greater stability in their decision patterns. A qualitative analysis of agents' deliberation transcripts indicates that these biases emerge from algorithmic shortcuts triggered under conditions of ambiguity and informational scarcity. In addition, characteristics of the assigned personas show significant correlations with the resulting ATEs, suggesting post-treatment confounding. This confounding appears to reflect a "persona drift" problem, whereby linguistic cues embedded in experimental treatments reshape the internal representation of personas—ultimately compromising the validity of randomization.

This study makes both substantive and methodological contributions to research on decision-making and to the growing intersection of artificial intelligence and the social sciences. Substantively, it offers a systematic comparison between human and LLM-based foreign policy decision-making, highlighting the epistemic risks of integrating LLMs into

political and security decision processes or into any decision environment characterized by comparable complexity and uncertainty. Methodologically, the study underscores the need for a careful rethinking of validity, randomization, and internal consistency when experiments incorporate generative AI agents. More broadly, it advances an agenda for using LLMs as instruments for theory testing and behavioral modeling, while delineating the epistemological and methodological boundaries necessary for their responsible application in social science research.

The remainder of the paper is organized as follows. Section 2 reviews key cognitive biases identified in human and machine decision-making, with particular attention to the characteristics of foreign policy decision environments and their implications. Section 3 outlines the research design and explains how the original human-subject experiments are replicated using LLM-based agents. Section 4 presents the empirical results, comparative analyses, and additional tests to interpret the findings. The final section discusses the broader theoretical and methodological implications of this research, its limitations, and directions for future work.

2 Biases in Human and Machine Decision-Making

International Relations (IR) scholars have long recognized that foreign policy decision-making is highly susceptible to cognitive and psychological biases. These biases often steer decisions toward a hawkish orientation, encouraging risk-seeking behavior, heightening the likelihood of conflict escalation, and reducing the willingness to compromise. Although group decision-making is frequently celebrated in other domains for fostering "collective wisdom," research in foreign policy suggests that it does little to correct such biases. On the contrary, group deliberation and collective processes may reinforce groupthink and amplify individual predispositions. As the world enters the AI era, a pressing and unresolved question emerges: can AI models and multi-agent systems perform better than humans under these conditions?

2.1 Human Cognitive Bias in Foreign Policy Decision-Making

Decision-making lies at the heart of politics, and there is a widespread expectation that high-stakes political choices—particularly in foreign policy—should rest on adequate information, clear preferences, and utility-maximizing calculations. Yet empirical research shows that such decision-making is deeply affected by cognitive biases (Jervis, 1976; Levy, 1997; Tetlock, 2005; McDermott, 2004; Vertzberger, 1998; Hafner-Burton, Hughes and Victor, 2013). This tendency is closely tied to the nature of foreign policy decisions, which typically unfold under conditions of profound uncertainty, information scarcity, and acute tension. These features make purely rational calculation difficult to sustain, and decision-makers inevitably rely on cognitive shortcuts, heuristics, dispositions, and even impulse to reach timely judgments.

Psychology has identified a wide array of cognitive biases, but those that have drawn particular attention in foreign policy are the biases that systematically push decision-makers toward confrontation. These "hawkish" tendencies involve overestimating threats, imputing hostile intent to adversaries, and favoring aggressive military or diplomatic responses (Jervis, 1976; Janis, 1982; Kahneman and Tversky, 1979; McDermott, 2004; Kertzer et al., 2022).

One of the central examples comes from prospect theory, which posits that decision-makers become unusually risk-seeking when facing potential losses—willing to accept risks that far exceed those they would tolerate in equivalent gain contexts. Prospect theory also predicts a framing effect: presenting the same situation in terms of losses rather than gains increases individuals' propensity to choose riskier options (Kahneman and Tversky, 1979; Levy, 1997; McDermott, 2004; Kertzer, Rathbun and Rathbun, 2020; Passarelli and Ponte, 2020).

Another component of the broader "hawkish bias" is the well-documented intentionality bias. This bias refers to the tendency of decision-makers to interpret adverse events as deliberate acts of hostility rather than as accidents, misperceptions, or consequences of situational constraints (Jervis, 1976; Kahneman and Renshon, 2007; Mercer, 2010; Chu, Holmes and Traven, 2021). Such interpretations can fuel unnecessary retaliation and escalation in

international politics, even when the triggering actions were unintended.

A third example is reactive devaluation, the tendency to systematically discount proposals that originate from an opponent or rival (Ross, 1993; Maoz et al., 2002). This bias makes conflict resolution through negotiation especially difficult, not because workable solutions are absent, but because mutual mistrust diminishes their perceived value.

Most studies of cognitive biases examine individual decision-making. Yet foreign policy choices are frequently made in group settings. Even when a decision is attributed to a single leader, it typically reflects prior consultation with advisors or members of the leader's inner circle. In many cases, decisions arise directly from collective deliberation and consensus among peers. This raises a central question: do collective decision-making and group deliberation reduce, sustain, or amplify cognitive biases in foreign policy decision-making?

The claim that groups can reduce biases originates from believe in the "wisdom of crowds"(George, 1972; Surowiecki, 2004). Studies suggest that in pluralistic groups, individuals' errors occur in different directions and degrees; when aggregated, these errors cancel out, producing wiser collective outcomes . In contrast, other studies find that groups often fail to improve decision quality and may even exacerbate cognitive pathologies. Group decision-making is as much a socio-psychological process as a rational one. In his seminal Groupthink, Janis (1982) shows that the pursuit of cohesion and harmony can override rational deliberation, amplifying bias rather than correcting it. Likewise, Mintz and Wayne (2016) identify the "Polythink" syndrome that fragmented groups whose lack of consensus undermines effective decision-making.

Recent experimental evidence supports the view that collective decision-making does not mitigate the hawkish bias in foreign policy. Kertzer et al. (2022) show that neither hierarchical nor horizontal groups significantly reduce hawkish biases compared to individual decision-making. Wayne et al. (2024) indicate that groups tend to over-estimate adversaries' resolve than individuals and are less likely to update their assessments when receiving new information.

In summary, humans are highly susceptible to cognitive biases and tend to make more "hawkish" choices under conditions of uncertainty, limited information, and heightened tension. Foreign policy elites are no exception: they exhibit the same psychological tendencies. Moreover, the introduction of group decision-making does not necessarily improve outcomes, as collective judgments may fail to outperform those of individuals.

2.2 AI as Decision-Maker: Better or Worse?

Given the pervasiveness of human biases in foreign policy decision-making, do AI systems reduce, amplify, or reproduce hawkish biases in individual or collective foreign policy decision-making?

To investigate this question, it is analytically useful to consider an extreme case–fully autonomous decision-making by machines. AI involvement in foreign policy could in principle span a broad spectrum, from data-assisted analysis and predictive modeling to semi-automated recommendations and, at the outer edge, autonomous execution of decisions. Although full automation remains speculative and distant, it provides a valuable conceptual benchmark: by exploring decision-making in the absence of human input, we can better identify machine-specific biases and distinguish them from those from humans.

Importantly, the application of AI in foreign policy decision-making is unlikely to proceed linearly along this automation spectrum. States will instead adopt and integrate AI systems selectively, depending on perceived utility, risk tolerance, institutional capacity, and geopolitical situations. Thus, examining the risks of fully autonomous decision-making is not a purely futuristic exercise; rather, it provides a necessary benchmark for evaluating how incremental forms of AI integration may already be reshaping biases, accountability, and the balance between human judgment and machine reasoning in international politics.

The appeal of delegating aspects of foreign policy decision-making to LLMs rests partly on the expectation that they approximate the ideal of rational actors more closely than human decision-makers. First, machines lack intrinsic cognitive worlds, emotional pressures,

and self-interests, allowing them to behave with greater internal consistency and rational coherence in high-stakes decision environments.[2] Second, because LLMs are trained on vast corpora, they possess extensive internalized knowledge that enables them to rapidly process complex and unstructured information, perform multi-step reasoning, and generate context-sensitive outputs.[3]

Moreover, recent advances in MASs allow for the construction of artificial societies in which agents assume specialized roles and interact strategically.[4] These LLM-based collectives can, in principle, approximate the "wisdom of crowds" without being vulnerable to emotional contagion, conformity pressures, or other social pathologies that frequently distort human group decisions.[5]

However, accumulating evidence shows that LLMs are far from immune to bias. Because these models are trained on human-generated data, they inevitably inherit the cognitive, cultural, and linguistic biases embedded in their training corpora. A growing body of work in cognitive science and AI alignment demonstrates that LLMs systematically reproduce human cognitive biases, mirroring the heuristics and error patterns that characterize human reasoning. Controlled experiments, for instance, reveal that LLMs display anchoring, availability, framing, and confirmation biases in probabilistic reasoning and moral judgment tasks.[6] LLMs also overweight initial prompts and contextual cues, favor information consistent with earlier statements, and exhibit order effects closely resembling those documented among human participants in behavioral experiments.[7]

[2]Simon, H. A. (1972). Theories of Bounded Rationality; Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

[3]OpenAI (2023). GPT-4 Technical Report; Bubeck, S. et al. (2023). Sparks of Artificial General Intelligence: Early Experiments with GPT-4. Microsoft Research.

[4]Park, J. S., O'Brien, J., Cai, C. J., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior; Wu, Z. et al. (2024). LLM-based Multi-Agent Simulation for Social Behavior Research.

[5]Janis, I. L. (1972). Victims of Groupthink; Sunstein, C. R., & Hastie, R. (2015). Wiser: Getting Beyond Groupthink to Make Groups Smarter. Harvard Business Review Press.

[6]Binz, M., & Schulz, E. (2023). Cognitive biases in large language models mirror those of humans. Nature Human Behaviour, 7, 1391–1401; Burton, J. W., Stein, M. K., & Jensen, T. B. (2024). Reproducing human biases in LLM reasoning tasks. Computational Cognitive Science.

[7]Zheng, K., et al. (2024). Anchoring and Framing Effects in LLM Decision-Making. Nature Machine Intelligence.

Further research indicates that reinforcement learning from human feedback (RLHF) may amplify these distortions, as the feedback guiding model optimization often reflects prevailing social, cultural, and ideological preferences (Chen, Hu and Lu, 2025; Malberg et al., 2024; Suri et al., 2023). In this sense, LLMs do more than mimic human language– they internalize and reproduce the patterned irrationalities that shape human judgment and decision-making.

Beyond cognitive biases, LLMs also exhibit systematic data- and model-level distortions. One prominent form is deviation bias, referring to systematic discrepancies between model outputs and real-world demographic or linguistic distributions. This stems largely from sampling imbalances and asymmetric optimization objectives, which amplify the representational dominance of majority-group data while marginalizing minority-group signals (Wang et al., 2025). Relatedly, social-category and demographic biases emerge when models generate disproportionately negative or stereotyped associations with specific identities, such as ethnicity, gender, or disability, despite mitigation efforts through prompt engineering and fine-tuning.[8]

When considering biases that could affect foreign policy, LLM's geopolitical bias represents another critical dimension. Because training corpora are linguistically and culturally bounded, LLMs often reproduce the narrative framings and ideological orientations of their data environments. For example, models trained predominantly on English-language or Western sources tend to portray international issues through the lens of liberal institutionalism, while models trained in other linguistic ecosystems may mirror their domestic political narratives (Salnikov et al., 2025) .

MASs may be no less susceptible to such biases. Emerging analyses show that LLM-based multi-agent systems, through natural-language interaction, can spontaneously develop human-like linguistic norms and collective biases. In simulated social environments, agents gradually converge on shared modes of expression through repeated interaction, forming vir-

[8](Gupta et al., 2025); Liu, S. (2025). Evaluating Bias Mitigation Strategies in Large Language Models.Zhao, Q. (2025). Explicit and Implicit Social Bias in LLM Outputs.

tual social conventions. Even when individual agents begin unbiased, group-level dynamics can lead the collective to privilege certain conventions, thereby entrenching biased traditions as emergent linguistic norms (Ashery, Aiello and Baronchelli, 2025).

These processes are highly sensitive to communication structures, as network topology shapes how information flows and influence accumulates. They are also deeply influenced by institutional design, such as governance rules, role assignments, and socioeconomic principles—which together determine agents' decision boundaries and structure the interaction patterns that ultimately shape collective outcomes (Jin et al., 2025; Zhuge et al., 2023).

3 Research Design

In this study, we examine how state-of-the-art LLMs make foreign policy decisions by benchmarking their choices against those of human participants. Specifically, we focus on the manifestation of hawkish bias in foreign policy decision-making, analyzing both individual (single-agent) choices and collective outcomes generated by multi-agent systems. This section presents the benchmark study and details the design of our replication experiments.

3.1 Benchmark Research and Overview of Experiment Design

Our benchmark study is Kertzer et al. (2022), which employed randomized controlled trials (RCTs) and effectively recruited 3,970 participants from an online experimental platform. The study investigates three well-documented biases in foreign policy decision-making, namely the loss framing bias, the intentionality bias, and the bias of reactive devaluation. Table 1 summarizes the scenario, treatment design, and dependent variable (decision to be made) for each experiment. Refer to the original study for more details.

The experiments were conducted under three conditions: individual, hierarchical group, and horizontal group. After completing demographic questionnaires individually, participants were randomly assigned to one of these conditions. In the individual condition,

Table 1: Original Experimental Design of Kertzer et al. (2022)

Bias	Scenario	Treatment	Control	Choice
Prospect theory	In a war-torn region, the lives of 600 stranded people are at stake. Two response plans with the following potential outcomes have been proposed by your advisors:	Loss frame-Policy A: 400 people will die; Policy B: There is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die	Gain frame-Policy A: 200 people will be saved; Policy B: There is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved	Policy A or B
Intentionality bias	Suppose that you are US policy-makers working on the North Korea conflict. You have just received a report that a US navy vessel has sunk 100 miles northeast of North Korean shores.	Fatalities-Unfortunately, there were 100 fatalities as none of the service people on the boat could be rescued.	No fatalities-Fortunately, there were no fatalities as all service people on the boat were rescued.	How likely did you think it was that the vessel was intentionally sunk? (1-7)
Reactive devaluation	Recently, the United States and Chinese governments held low-level talks with the aim of trying to resolve ongoing disputes over trade. Last week, the (Chinese/US) government submitted a brief proposal...	China authorship	US authorship	How much do you support the proposal, on a scale from 1 to 7?

participants made decisions on various foreign policy options independently. In the group conditions, participants were randomly assigned to groups of 3–5 members and engaged in deliberation via an online chatroom before reaching a decision. In horizontal groups, all members had equal status and were required to reach a collective, unanimous decision. In hierarchical groups, one member was randomly designated as the leader with final decision-making authority, while the remaining members acted as advisors, providing counsel to the leader. Figure 1 summarizes the experimental procedure.

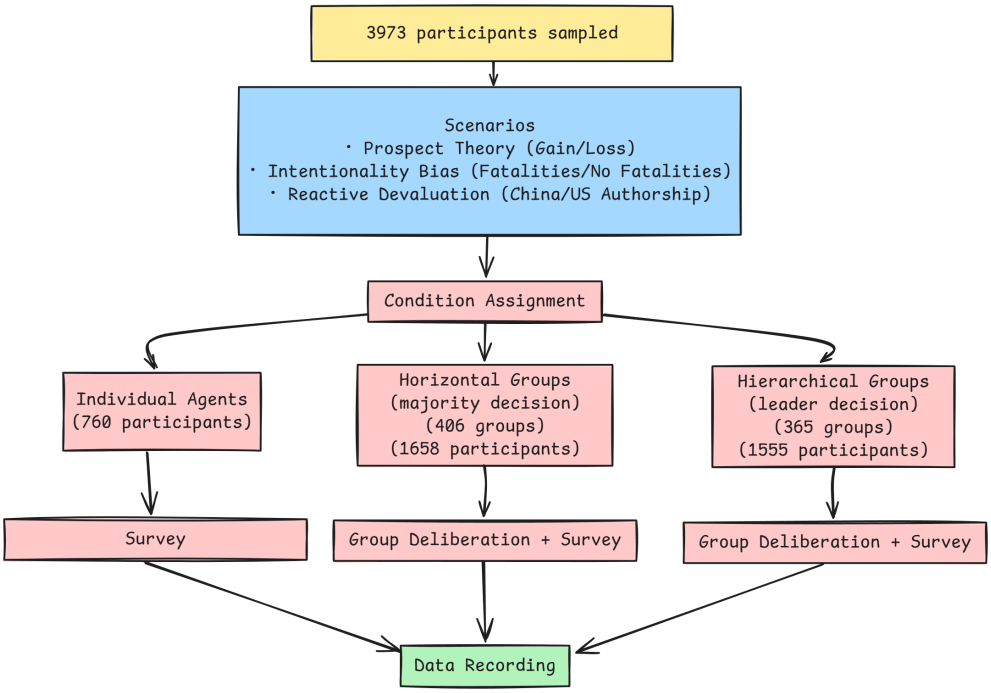

Figure 1: Experimental Flowchart

Kertzer et al. (2022) provides a valuable benchmark for our study for several reasons. First, randomization ensures that observed differences in choices can be attributed to treatment effects (i.e., biases) rather than uncontrolled variation. Second, the stylized design of the experiments enhances comparability between human participants and LLMs: in simplified, well-defined settings, LLM responses can be systematically evaluated against human behavior under identical conditions. Third, because the original experiment employs textual prompts as treatments, it aligns closely with the linguistic foundations of LLMs, which

process and respond to textual cues in ways analogous to humans. Finally, the experiments aim not to assess correctness but to reveal systematic tendencies, precisely the behavioral deviations that constitute "bias" in foreign policy decision-making.

3.2 LLM Replication Framework

Our replication is based on the same sample size, employing 3,973 synthetic participants. We strictly follow the experiment design and implementation as in the original study. The agents are distributed across experimental conditions following the original study's proportions: 760 individual agents, 406 horizontal groups (1,658 participants), and 365 hierarchical groups (1,555 participants). Each synthetic participant is assigned to all three experimental scenarios with randomized treatment assignment within each scenario.

To conduct a synthesized group decision-making to mimic an chatroom environment, we build a multi-agent environment, as illustrated in Figure 2. This system comprises the methodology for prompt setting (left) and the specific experimental workflow of LLM agents (right). The prompt engineering framework consists of four key components. The first is role settings, where LLM-synthesized participants are instructed to act as expert role-players. The second component is the experimental scenario, determined by the specific experiment and randomized treatment allocation, with corresponding text references provided in Table 1. The third component is the persona profile, which includes three parts: demographic information (age, gender, income, education), beliefs (covering sociopolitical attitudes such as militant internationalism, isolationism, risk acceptance, social dominance orientation and so on, as well as personal traits like the Big Five personality dimensions), and an AI-generated narrative integrating these features cohesively. The persona construction methodology follows recent advances in population seeding (Park et al., 2024; Chuang et al., 2024), which demonstrate that belief-related attributes significantly increase individual-level prediction accuracy. The fourth component is task information, containing rules governing interactions between synthetic agents, scenario details, and instructions distinguishing

different roles across conditions (e.g., leader, advisor). After providing these prompts, we present the question to the LLM agent and request its response. We report in the appendix the technical details of constructing agents and multi-agent environment.

The experimental flow demonstrates how all synthesized participants are systematically assigned to these parallel conditions. Participants in the horizontal group (1,658 participants, 406 groups) engage in multi-agent deliberation based on the premise of member equality and are required to reach a consensus and unanimous decision (if consensus cannot be achieved, we ask each group member for their individual choice and aggregate the group decision using either the median or majority rule). Participants in the hierarchical group (1,555 participants, 365 groups) are led by a randomly assigned leader who makes the final decision after deliberation with the remaining four advisors. In all groups, deliberation unfolds in iterative rounds. We create an automated supervisor to control the progress of deliberation (including rounds, timing, and actions). When the supervisor permits agents to begin deliberation, in each round, all group members contribute their content in a chatbox for the supervisor to manage rounds and progress. We also generate a summarizer agent to periodically compile the discussion progress. Once deliberation concludes and the summary is complete, all agents undergo a final survey regarding the dependent variable (the decision) and their confidence in their decision. Note that within groups, regardless of whether participants have substantive influence on the decision, we ask each participant for their preferred decision choice and confidence level.

4 Empirical Findings

We perform the synthetic replication and analyze the resulting data. This section presents and interprets the evidence. In addition, to contextualize the replication outcomes, we qualitatively examine the deliberation scripts produced by MASs and highlight some initial insights.

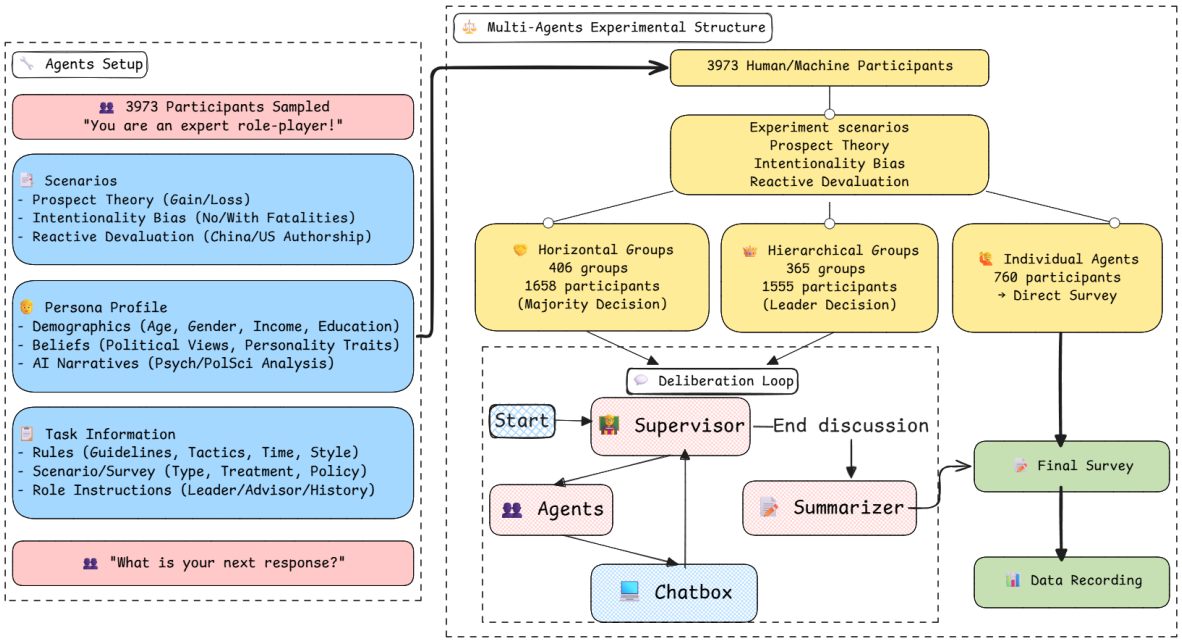

Figure 2: Replication Pipeline

4.1 Average Treatment Effects in Synthetic Replications

We begin by comparing the average treatment effects (ATEs) between our synthetic replications and the original experiments.

4.1.1 The Loss Framing Bias

Figure 3 illustrates the effects of the loss frame on the probability of risky decisions in the prospect theory scenario, showing results for both LLM agents and human participants across different group contexts. Consistent with the original findings, the loss frame significantly increases risk-taking for both humans and LLMs. In the individual context, LLM agents exhibit a stronger risk preference under the loss frame than human participants. However, MASs do not uniformly mitigate group pathologies: while the effect of the loss frame decreases in hierarchical group settings relative to the individual context, it remains largely unchanged in horizontal group settings. Overall, these results indicate that LLMs replicate human biases in the domain of loss and language framing, though group dynamics can modulate—but do not eliminate—these tendencies.

Figure 4 further illustrates the effects of the loss frame on LLM agents and human participants under different horizontal decision rules. Across all three rules–majority, median voter, and unanimity–the pattern of bias (ATE) is similar between LLMs and humans, although the effect is generally larger in the synthetic replication. Notably, the loss frame effect is strongest under the unanimity rule compared to the other decision rules.

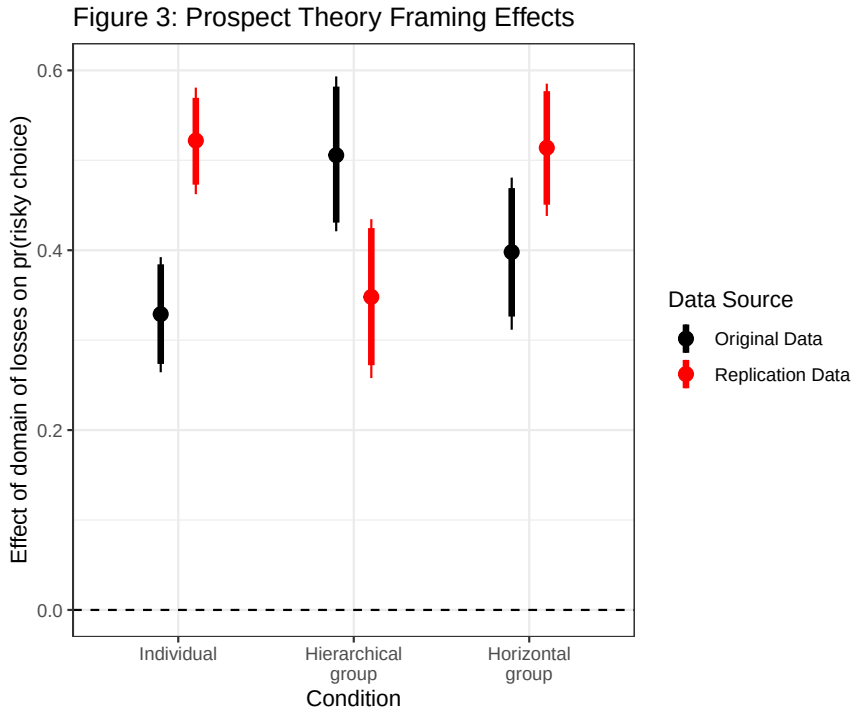

Figure 3: Comparison of effects: Prospect Theory Scenario
Note: In this figure, black points and bars represent the original data from human participants, while red points and bars represent the replication data from LLM agents. Horizontal group decisions are calculated using the median voter aggregation method. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.

4.1.2 The Intentionality Bias

Figure 5 presents the ATEs in the intentionality bias scenario. In contrast to the prospect theory setting, LLM agents largely reverse the pattern of human biases, exhibiting a distinct "dovish bias." The point estimates of ATEs in the synthetic replications are negative–opposite in direction to those observed in human experiments. According to the original findings by Kertzer et al. (2022), when presented with information involving fatalities, hu-

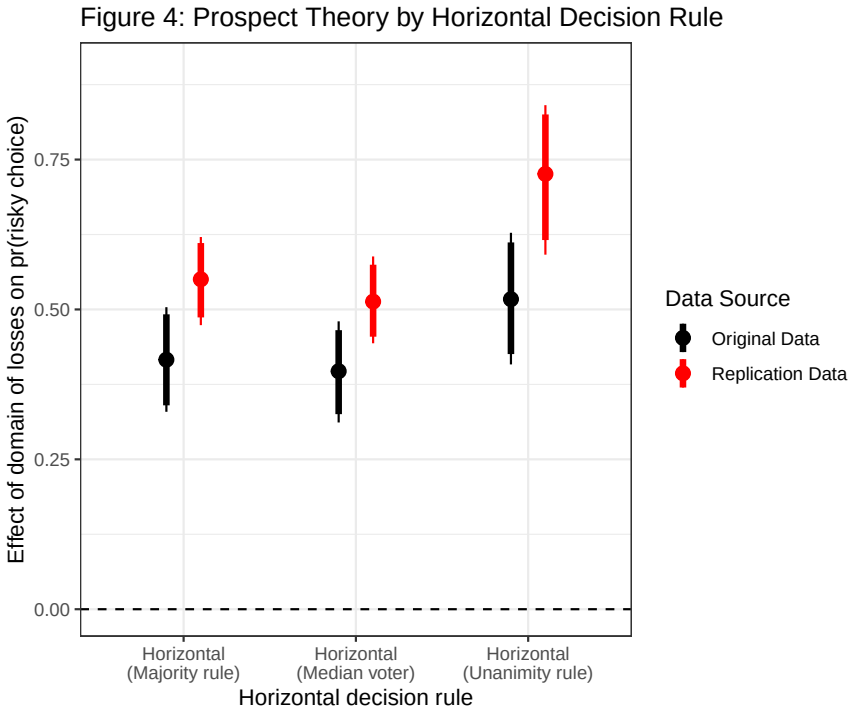

Figure 4: Comparison of effects: Prospect Theory Scenario (by Different Horizontal Decision Rules)
Note: In this figure, black points and bars represent the original data from human participants, while red points and pars represent the replication data from LLM agents. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.

man participants are more likely to interpret the sinking of a ship as an intentional act by North Korea. LLM agents, however, display the opposite tendency: exposure to information without fatalities makes them more likely to infer intentionality. This divergence suggests that the causal mechanisms underlying LLM and human decision-making in the intentionality bias scenario differ fundamentally. Machines thus reveal a novel form of bias that departs from known human behavioral patterns.

It is worth noting that the ATE for horizontal groups with LLM participants is not statistically significant. We further examine whether different decision rules alter the collective outcome. As shown in Figure 6, the aggregation rule indeed affects machine group decision-making. While the median voter and majority rules effectively mitigate bias, the unanimity rule produces an ATE closely resembling that of human participants.

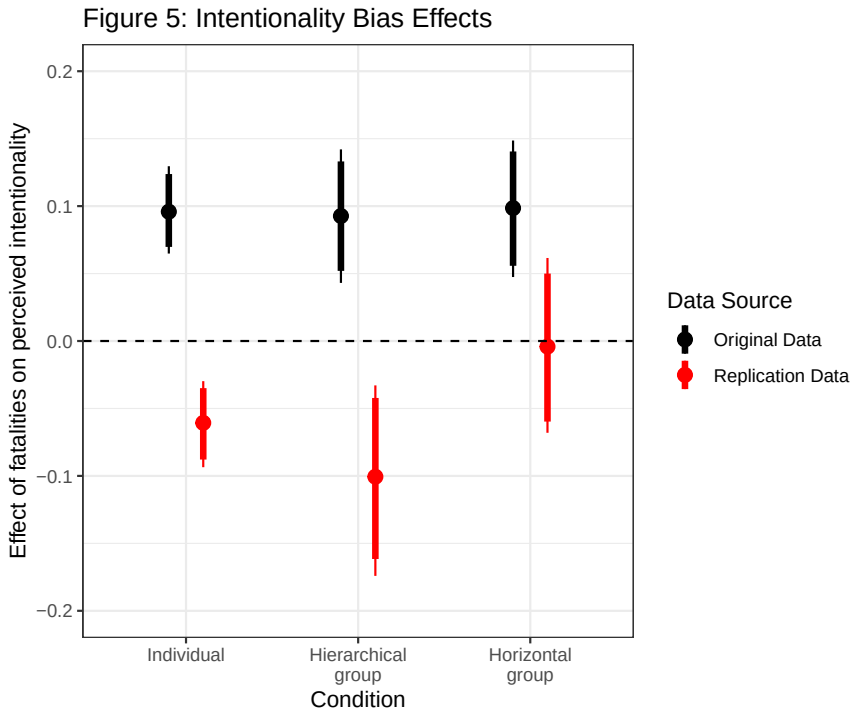

Figure 5: Comparison of effects: Intentionality Bias Scenario
Note: In this figure, black points and bars represent the original data from human subjects, while red points represent the replication data from LLM agents. Horizontal group decisions are calculated using the median voter aggregation method. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.

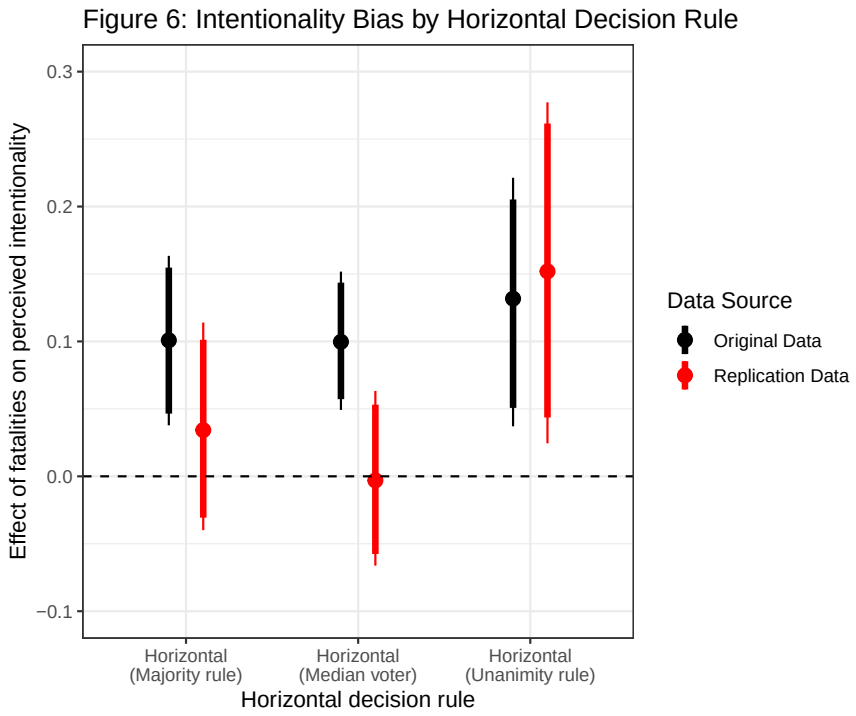

Figure 6: Comparison of effects: Intentionality Bias Scenario (by Different Horizontal Decision Rules)
Note: In this figure, black points and bars represent the original data from human participants, while red points and pars represent the replication data from LLM agents. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.

4.1.3 The Reactive Devaluation Bias

In the reactive devaluation scenario, LLM agents exhibit more consistent biases than human participants. Figure 7 displays the effect of China authorship on support for an agreement for both groups. Among human participants, the direction and significance of the coefficients are mixed—only the horizontal groups show the expected bias. In contrast, the point estimates of ATEs in the synthetic replication consistently demonstrate reactive devaluation: across individual, hierarchical, and horizontal contexts, attributing the agreement to China decreases their level of support.

As shown in Figure 8, voting rules again appear to influence outcomes. However, the patterns diverge between humans and machines. In the human experiments, horizontal decision-making under various rules produces the most robust bias, whereas for LLMs, unanimity and majority rules effectively mitigate the bias. This finding contrasts with the intentionality experiment, where unanimity amplified machine bias, making it difficult to draw a general conclusion about which aggregation rule best mitigates systematic biases.

In Kertzer et al. (2022), they interpret the inconsistency in human experimental results as stemming from the high detailedness of the proposal, which mitigate the mechanism of reactive devaluation they aim to test-that proposals from an opponent are more likely to be questioned due to perceived neglect of national interests. The overdetailed proposal distract human participants. However, LLMs possess a stronger ability than humans to precisely extract relevant information as prompted. In the deliberation scripts, many LLM agents indeed notice that China is the author of the agreement and begin reflecting on the reliability of the proposed agreement based on the history of Sino-U.S. relations. In this sense, LLMs manifest human-like biases.

4.1.4 Does Deliberation Matter?

Both the human experiments and their synthetic replications suggest that institutional design–specifically, the distribution of authority and the aggregation rule–plays a mean-

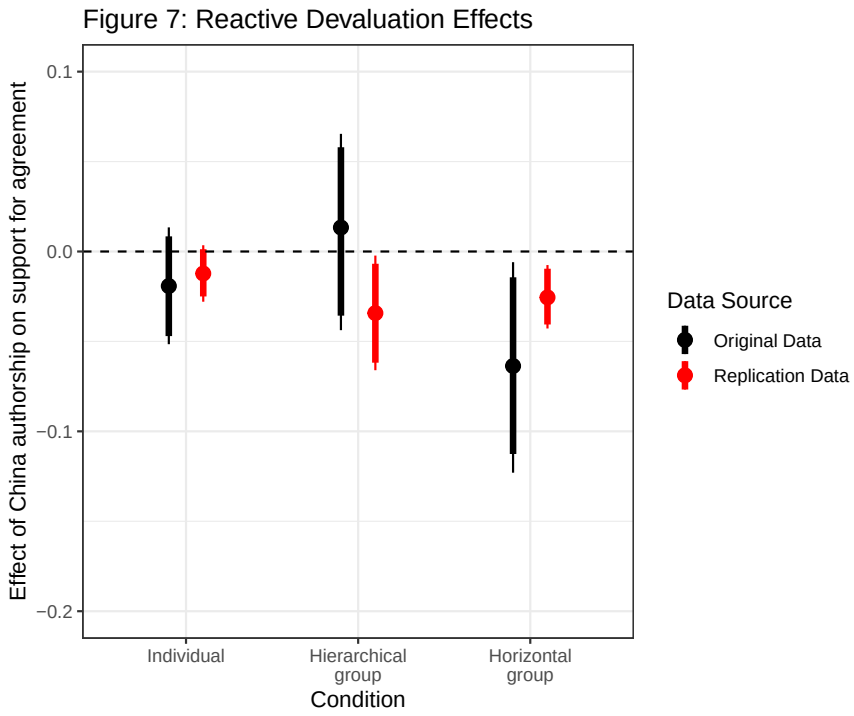

Figure 7: Comparison of effects: Reactive Devaluation Scenario

Note: In this figure, black points and bars represent the original data from human participants, while red points and bars represent the replication data from LLM agents. Horizontal group decisions are calculated using the median voter aggregation method. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.

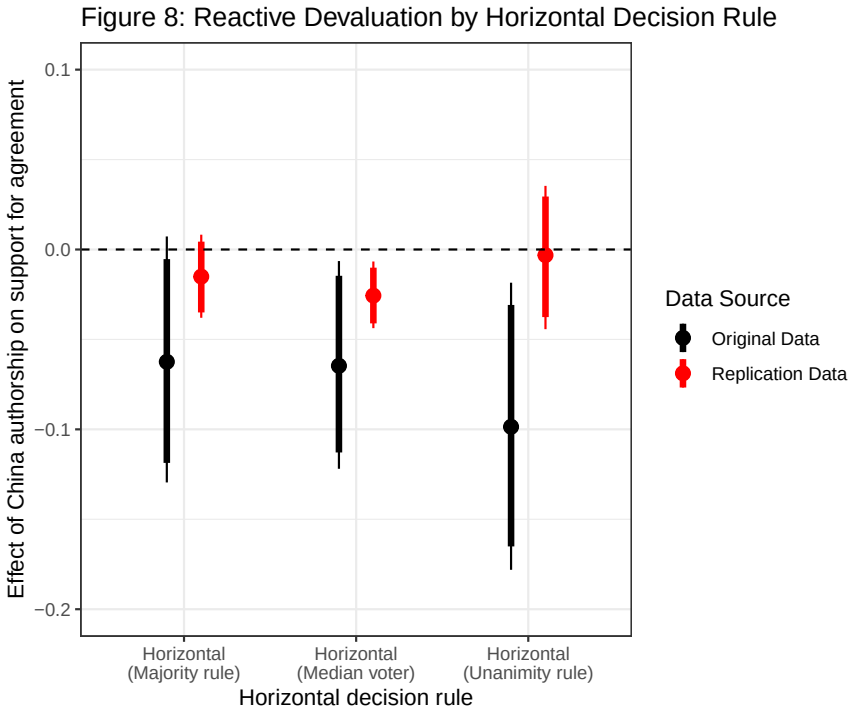

Figure 8: Comparison of effects: Reactive Devaluation Scenario (by Different Horizontal Decision Rules)
Note: In this figure, black points and bars represent the original data from human participants, while red points and pars represent the replication data from LLM agents. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.

ingful role in shaping outcomes. However, shifting attention from outcomes to processes raises another important question: can group deliberation itself alter individual preferences? In our design, each MAS records the preferred choices of its members after the group decision is made. This enables a comparison between the distribution of individual choices in the individual decision condition and those in the group decision conditions. Since LLM agents are randomly assigned to different decision contexts, any systematic differences between these distributions can be attributed to the deliberative process.

Figure Figure 9 summarizes the distributions from the original and synthetic experiments. In the prospect theory scenario, deliberation does not alter the proportion of human subjects choosing the risky option, whereas among LLM agents, deliberation appears to reduce the likelihood of selecting the risky policy B. In the intentionality bias scenario, deliberation has heterogeneous effects: within groups, humans tend to become more hawkish, while LLMs shift toward more dovish positions. Among the three experimental settings, intentionality bias is the most urgent and information-scarce, features that were repeatedly emphasized during LLM group discussions and ultimately led to a consensus favoring conservative and cautious conclusions. In the reactive devaluation scenario, LLM choices are more concentrated than those of humans, regardless of group context, and deliberation does not affect this concentration.

It is difficult to draw a conclusion about whether, when, and how deliberation influences the hawkish bias. The experimental results suggest that the effect of deliberation is context-dependent and may vary across informational environments and cognitive frames, rather than exerting a uniform moderating or amplifying influence on hawkish tendencies.

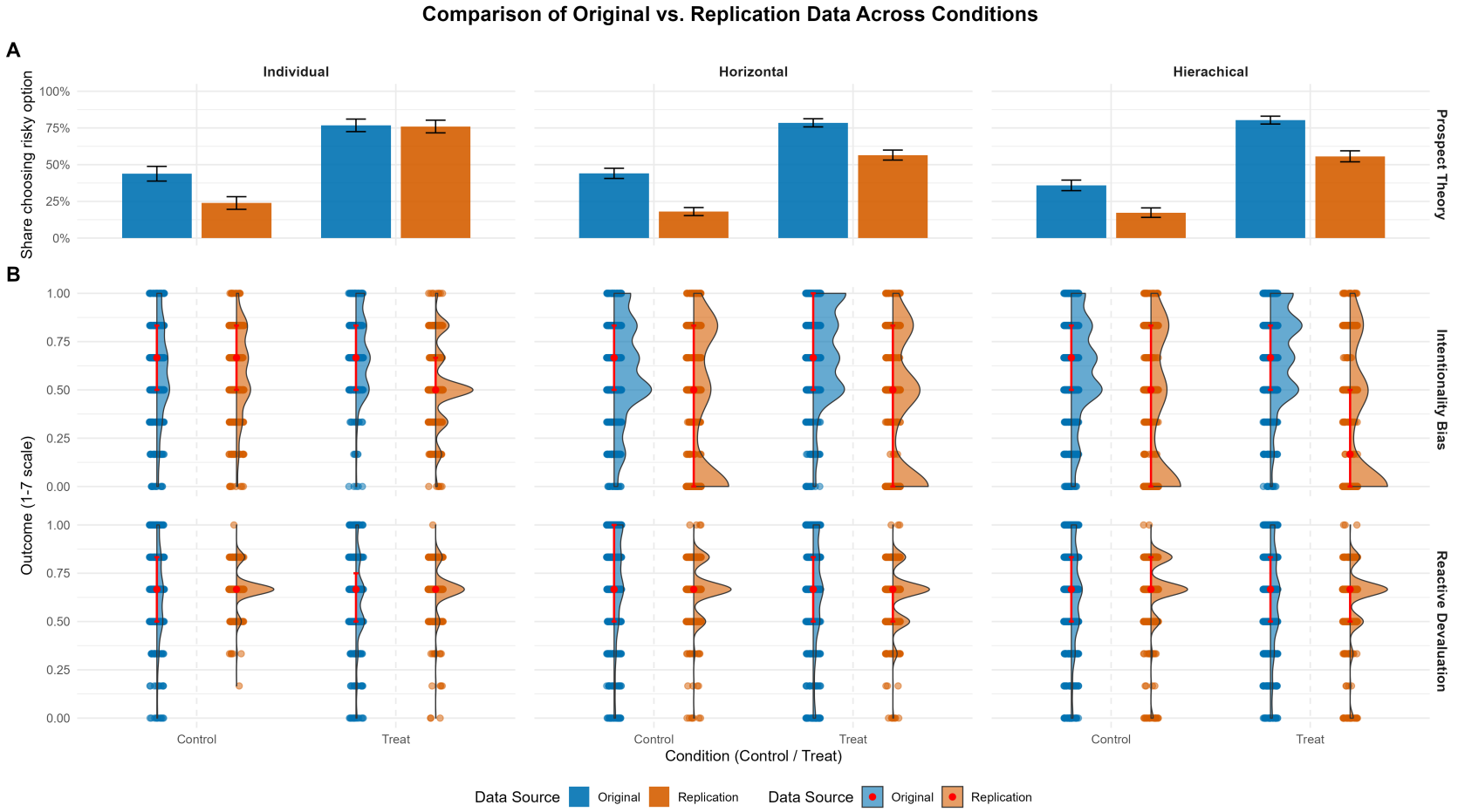

Figure 9: Comparison of Distribution between LLM Agents and Human Participants across Scenarios and Group Contexts

We can summarize the main findings as follows. First, in foreign policy decision-making, machines are no less susceptible than humans to cognitive biases, although the patterns of these biases differ substantially. Second, MAS (deliberative groups composed of LLMs in this paper) do not exhibit reduced bias compared to humans or single-agent models; only under conditions of urgency and information scarcity do MAS deliberations tend to shift decisions toward more dovish positions. Third, our synthetic replications reveal that the biases displayed by LLMs are far from simple reproductions of human biases. While LLMs replicate human tendencies under prospect theory conditions, they diverge notably in scenarios involving intentionality bias and reactive devaluation. These inconsistent patterns make machine biases difficult for humans to anticipate or interpret.

4.2 Understanding the Patterns of Biases in LLM's Decision

A central question is, given that LLMs, as machines, naturally lack the inherent human "cognitive world," where do their "cognitive biases" originate? This question is difficult to answer since LLMs are essentially blackboxes.

Machines, of course, differ fundamentally from humans and possess their own types of shortcuts when confronted with high uncertainty, complexity, or incomplete information—shortcuts that can themselves produce biases. For humans, the ideal model of decision-making assumes complete information, certainty, and rational utility maximization. When these conditions are absent, humans rely on cognitive heuristics shaped by evolutionary adaptation, individual experience, and sociocultural context.

For machines, the logic is different. Designed to follow principles of efficiency, optimization, and iterative updating, machines face bias not through affective or experiential shortcuts but through algorithmic ones. Under conditions of uncertainty or incomplete information—where analytical solutions cannot be precisely derived—machines resort to heuristic or approximate algorithms to generate answers. Consequently, humans and machines are not only sensitive to different cues, but may also respond in fundamentally different ways to

the same cue. Understanding machine bias therefore requires a careful analysis of text-as-treatment: identifying the linguistic cues and combinations that give rise to distinct forms of bias in LLM-driven decision-making.

An even more complex and consequential source of bias arises from persona assignment. Assigning personas to LLMs can introduce additional distortions linked to the well-documented phenomenon of persona drift, whereby the meaning and behavioral consistency of an assigned identity evolve unpredictably over time or across contexts. The rationale for assigning personas is to make LLM agents' behaviors more interpretable. However, LLMs may generate responses based on simplified or stereotypical representations of personas embedded in their training data, thereby introducing additional biases (Dash et al., 2025). More critically, the meaning of an assigned persona can itself change following treatment–a phenomenon known as persona drift. Data-driven responsive patterns could lead to unintended associations between persona cues and certain linguistic cues in the treatment, causing systematic deviations in LLM responses.In this case, it is not the values of pre-treatment variables (the persona profiles) that change, but rather the meaning of those variables. Such drift produces post-treatment confounding, making causal effects uninterpretable and the identified quantity inconsistent with the intended estimand.

Therefore, we regress the difference between LLM and human responses on all demographic and ideological attributes, and all the coefficients across three experimental scenarios and different AI settings are shown in Figure 10. If the LLM indeed simulates human cognitive patterns well, then the coefficient of any demographic or ideological attribute should not be significant–it should not significantly increase or decrease the distance between LLM and human responses.

However, the results show that some attributes significantly increase or reduce the distance, but these significant effects are not robust. First, the same attribute may positively predict the distance in one experimental scenario but be insignificant or even negatively signed in others, such as openness and right-wing authoritarianism. This indicates that

the LLM does not consistently associate specific outputs with attribute settings, but instead makes heterogeneous adjustments based on both attributes and experimental scenarios. Second, even within the same experimental scenario, the coefficient of the same attribute varies significantly depending on LLM settings, such as militant internationalism and isolationism. This suggests that the way agent attributes are set also moderates the accuracy with which LLMs simulate specific human attributes. These results at least indicate that persona cues provided to LLM agents introduce post-treatment confounding: assigning persona traits to LLMs (and assigning in different ways) interacts with specific cues in experimental scenarios, systematically biasing ATEs.

We also regress the response on interaction terms between all persona attributes and AI settings, with results shown in Figure 11. In human experiments, due to randomization, no pre-treatment attribute should significantly predict responses. However, in LLM agent experiments, even with randomized assignment, a considerable number of persona attributes remain significantly associated with the LLM agent's outputs. Similar to the observations in Figure 10, the significance and signs of these coefficients are not entirely consistent across different experimental scenarios and AI settings. In intentionality bias, need for cognition consistently and significantly negatively predicts the output (perceived intentionality attributed to North Korea) across various settings, likely because the cue "without further information" in the vignette text, when combined with need for cognition, triggers cautious and conservative evaluations. Overall, in the absence of exact analytical solutions for specific questions, LLMs systematically link predefined persona cues and provided vignette text cues to specific outputs through algorithmic heuristic shortcuts, which become post-treatment confounders in RCTs.

Evidence from the deliberation transcripts further supports the finding that LLMs rely on their assigned personas in ways correlated with the treatment, particularly when making decisions under conditions of information scarcity. Unlike humans, who tend to employ cognitive or affective shortcuts when faced with limited information, LLMs exhibit a

distinct information-seeking tendency. Under conditions of uncertainty and scarce information, LLMs often over-empower surface-level linguistic cues (such as a particular phrase, high-frequency vocabulary, or linguistic style), treating them as reliable signals (Bender and Koller, 2020); whereas humans tend to interpret and be emotionally influenced by the hidden deeper meanings behind these cues (Kahneman, Slovic and Tversky, 1982). In the intentionality bias experiments, nearly all LLM participants carefully avoided premature judgments, instead reaching a shared consensus to seek additional evidence before attributing hostile intent to North Korea. The phrase "without further information" frequently appeared as a salient linguistic cue prompting restraint. This pattern helps explain why fatalities did not increase attributions of hostile intent: whereas emotionally charged cues trigger affective responses in humans, LLM deliberations were dominated by a rationalized demand for more evidence. LLM-driven agents thus demonstrated high sensitivity to textual cues and a recurring tendency to draw on their assigned personas as a source of inferred information.

This dynamic is further illustrated throughout the transcripts. During deliberations, LLM agents consistently and explicitly referenced their personas, invoking attributes such as political affiliation, religious identity, or risk tolerance to justify their positions. Once such self-anchoring occurred, agents rarely reversed course. This suggests that LLM agents blur the distinction between treatment and pre-treatment conditions: lacking internalized and stable dispositions, they effectively have no genuine "pre-treatment" state. Consequently, assigning personas to LLMs prior to treatment is qualitatively different from recording pre-treatment characteristics in human subjects.

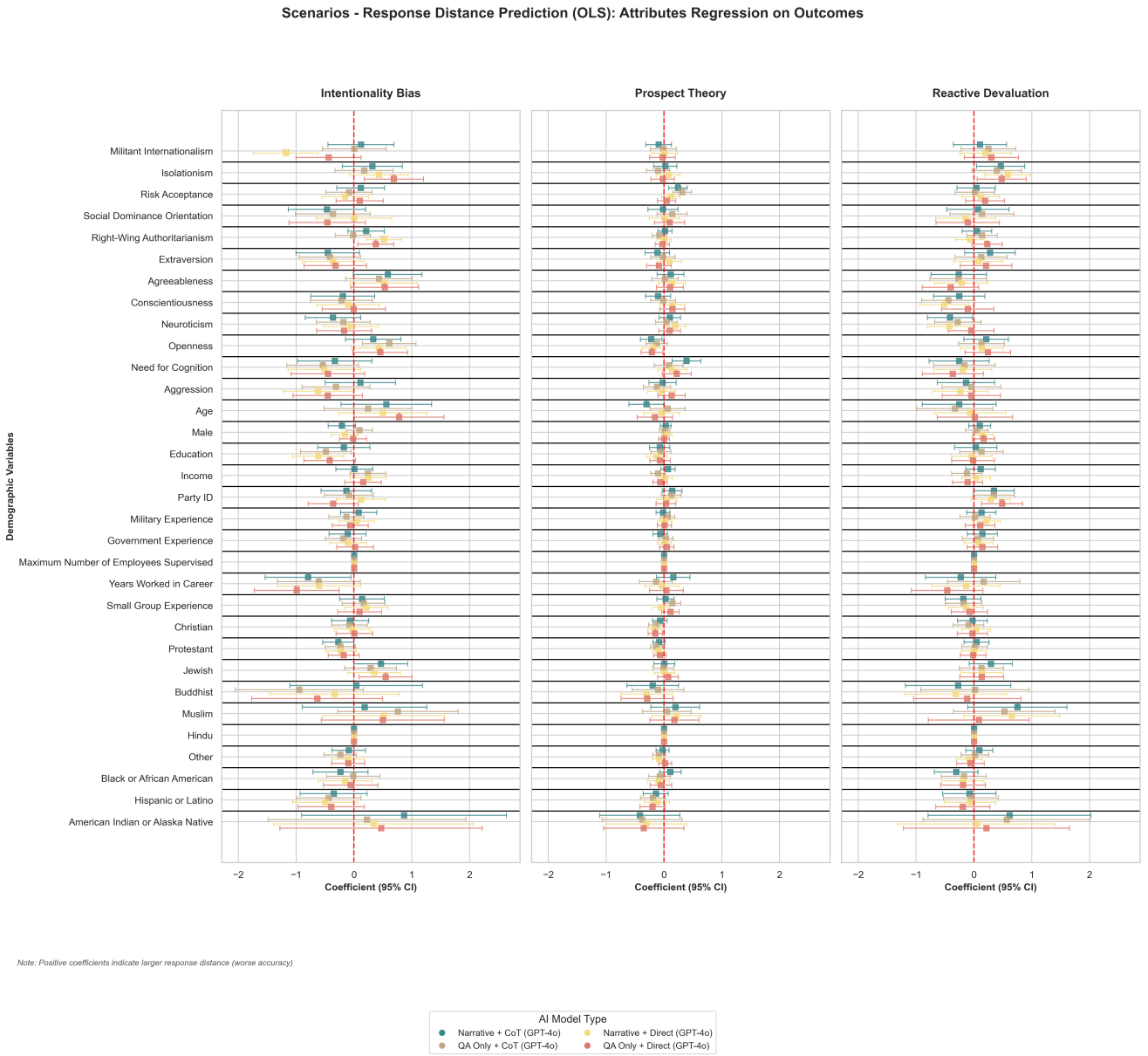

Figure 10: Regression Results of Demographic Variables on LLM-Human Response Distance

Figure 11: Regression Results of Demographic Attributes × AI Settings on Response

5 Conclusion and Discussion

In this study, we replicate experiments on hawkish bias in foreign policy decision-making (Kertzer et al., 2022) using LLM-based multi-agent systems (MAS). Our goal is to identify the biases exhibited by machine agents and compare them with known human biases, thereby assessing the risks of deploying LLMs in decision environments characterized by scarce information, unstructured inputs, and time-sensitive demands. We suspect that LLMs may be especially appealing to human decision-makers in such contexts because they can help offload moral, affective, and cognitive burdens associated with high-stakes choices.

A second central objective is to evaluate whether LLM-based agents are ready to serve as synthetic participants–either as substitutes for, or complements to, human-subject RCTs, which remain the gold standard for causal inference in the social sciences. This line of inquiry is particularly attractive because it promises to lower the cost and reduce the logistical and ethical constraints that currently limit many human-subject studies.

Our findings are pessimistic to both real-world and academic applications. We find that synthetic decision-makers exhibit systematic biases in these scenarios, and that MAS neither mitigate these biases nor make them more predictable. Although different aggregation rules do shift collective outcomes, they do not do so in consistent or theoretically expected ways. Deliberation similarly fails to make agents more "rational," either individually or collectively. More troublingly, machine biases diverge from human biases in ways that are both uninterpretable and unpredictable, making it unclear whether—and how—human–machine joint decision-making could be rendered more rational. These biases arise from shortcuts fundamentally different from human heuristics: LLMs rely on algorithmic heuristics and approximations because uncertainty, complexity, and incomplete information preclude precise analytical solutions. As a result, humans and machines respond not only to different cues but also in qualitatively different ways to the same cues.

These cognitive biases raise significant concerns for foreign policy, political, and public policy decision-makers. The opacity and instability of LLM biases make their deployment in

high-stakes, unstructured, ambiguous, and time-sensitive decision environments risky—even dangerous. MAS do not reliably correct human group pathologies and instead introduce machine-specific group pathologies, warranting caution before placing trust in the "wisdom of machine crowds."

For social scientists, our findings also deliver sobering implications for LLM-based experimental research. Randomized controlled trials that assign personas to LLM agents may unintentionally introduce post-treatment confounding because LLMs are highly sensitive to textual cues and prone to automatic associative reasoning. As a consequence, LLM-based agents cannot yet be considered reliable substitutes for, or supplements to, human subjects in social science experiments—particularly those that rely on text-as-treatment designs.

This study has several limitations that point to important directions for future research. First, our findings rely on a preliminary replication of benchmark experiments that themselves may contain design limitations. Future work should systematically test which components of these designs contribute to the emergence of machine biases, thereby improving interpretability. Second, machine biases may not be inherently unpredictable or uninterpretable; large-scale synthetic replications may reveal stable patterns that are not visible in smaller samples. Our results therefore call for more, not fewer, synthetic replications. Third, while we argue that current LLM-based synthetic participants are ill-suited for RCTs due to persona drift, this does not imply that the approach is unworkable. Instead, it highlights the need for methodological advances to address persona drift and to better understand how machine responses to linguistic cues differ from those of humans—work that is essential for developing a rigorous experimental framework for synthetic participants.

References

Ashery, Ariel Flint, Luca Maria Aiello and Andrea Baronchelli. 2025. "Emergent Social Conventions and Collective Bias in LLM Populations." *Science Advances* 11(20):eadu9368.

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ed. Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault. Online: Association for Computational Linguistics pp. 5185–5198.

Chen, Yaoyu, Yuheng Hu and Yingda Lu. 2025. "Predicting Field Experiments with Large Language Models.".

Chu, Jonathan A., Marcus Holmes and David Traven. 2021. "Inferring Intentions from Consequences: How Moral Judgments Shape Citizen Perceptions of Wartime Conduct." *Journal of Experimental Political Science* 8(2):203–207.

Chuang, Yun-Shiuan, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu and Timothy T. Rogers. 2024. "Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks.".

Dash, Saloni, Amélie Reymond, Emma S. Spiro and Aylin Caliskan. 2025. "Persona-Assigned Large Language Models Exhibit Human-Like Motivated Reasoning.".

George, Alexander L. 1972. "The Case for Multiple Advocacy in Making Foreign Policy." *The American Political Science Review* 66(3):751–785.

Gupta, Ojasvi, Stefano Marrone, Francesco Gargiulo, Rajesh Jaiswal and Lidia Marassi. 2025. "Understanding Social Biases in Large Language Models." *AI* 6(5):106.

Hafner-Burton, Emilie M., D. Alex Hughes and David G. Victor. 2013. "The Cognitive Revolution and the Political Psychology of Elite Decision Making." *Perspectives on Politics* 11(2):368–386.

Janis, Irving Lester. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. University Press.

Jin, Weiqiang, Hongyang Du, Biao Zhao, Xingwu Tian, Bohang Shi and Guang Yang. 2025. "A Comprehensive Survey on Multi-Agent Cooperative Decision-Making: Scenarios, Approaches, Challenges and Perspectives.".

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2):263–291.

Kahneman, Daniel and Jonathan Renshon. 2007. "Why Hawks Win." *Foreign Policy* (158):34–38.

Kahneman, Daniel, Paul Slovic and Amos Tversky, eds. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kertzer, Joshua D., Brian C. Rathbun and Nina Srinivasan Rathbun. 2020. "The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations." *International Organization* 74(1):95–118.

Kertzer, Joshua D., Marcus Holmes, Brad L. LeVeck and Carly Wayne. 2022. "Hawkish Biases and Group Decision Making." *International Organization* 76(3):513–548.

Levy, Jack S. 1997. "Prospect Theory, Rational Choice, and International Relations." *International Studies Quarterly* 41(1):87–112.

Malberg, Simon, Roman Poletukhin, Carolin M. Schuster and Georg Groh. 2024. "A Comprehensive Evaluation of Cognitive Biases in LLMs.".

Maoz, Ifat, Andrew Ward, Michael Katz and Lee Ross. 2002. "Reactive Devaluation of an "Israeli" vs. "Palestinian" Peace Proposal." *The Journal of Conflict Resolution* 46(4):515–546.

McDermott, Rose. 2004. "Prospect Theory in Political Science: Gains and Losses From the First Decade." *Political Psychology* 25(2):289–312.

Mercer, Jonathan. 2010. "Emotional Beliefs." *International Organization* 64(1):1–31.

Mintz, Alex and Carly Wayne. 2016. *The Polythink Syndrome: U.S. Foreign Policy Decisions on 9/11, Afghanistan, Iraq, Iran, Syria, and ISIS*. Stanford University Press.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang and Michael S. Bernstein. 2024. "Generative Agent Simulations of 1,000 People.".

Passarelli, Francesco and Alessandro Del Ponte. 2020. Prospect Theory, Loss Aversion, and Political Behavior. In *Oxford Research Encyclopedia of Politics*.

Ross, Lee. 1993. *Reactive Devaluation in Negotiation and Conflict Resolution*. Stanford Center on Conflict and Negotiation, Stanford University.

Salnikov, Mikhail, Dmitrii Korzh, Ivan Lazichny, Elvir Karimov, Artyom Iudin, Ivan Oseledets, Oleg Y. Rogov, Natalia Loukachevitch, Alexander Panchenko and Elena Tutubalina. 2025. "Geopolitical Biases in LLMs: What Are the "Good" and the "Bad" Countries According to Contemporary Language Models.".

Suri, Gaurav, Lily R. Slater, Ali Ziaee and Morgan Nguyen. 2023. "Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5.".

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. The

Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations New York, NY, US: Doubleday & Co.

Tetlock, Philip Eyrikson. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Vertzberger, Yaacov Y. I. 1998. *Risk Taking and Decision Making: Foreign Military Intervention Decisions*. Stanford University Press.

Wang, Daniel, Eli Brignac, Minjia Mao and Xiao Fang. 2025. "Measuring Stereotype and Deviation Biases in Large Language Models.".

Wayne, Carly, Mitsuru Mukaigawara, Joshua D. Kertzer and Marcus Holmes. 2024. "Diplomacy by Committee: Assessing Resolve and Costly Signals in Group Settings." *American Journal of Political Science* n/a(n/a).

Zhuge, Mingchen, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem and Jürgen Schmidhuber. 2023. "Mindstorms in Natural Language-Based Societies of Mind.".

Supplementary Material

A Multi-Agent System(MAS) Details

This appendix documents the technical details of our multi-agent large language model (LLM) framework and the implementation of the group deliberation experiments. The goal is to make the design transparent and replicable without overloading the main text with engineering choices. All code, configuration files, and raw outputs are provided in the replication materials.

To increase transparency while keeping exposition compact, Appendix Figures 12–13 summarize the full implementation as high-level pseudo-code, from persona calibration to LangGraph orchestration and estimation of experimental quantities.

A.1 Overall Framework and Environment

We implement the experiments using a graph-based workflow built on LangGraph. Rather than hard-coding a sequence of API calls, we represent a discussion as a state machine in which nodes correspond to functional modules (for example, environment initialization, coordination, individual agents, and chatbox processing) and directed edges specify how the discussion state flows between them. A single "discussion graph" orchestrates the entire process, including group-level control and individual agent behaviour, which ensures that horizontal and hierarchical groups, survey-only conditions, and different timing regimes are handled within a unified architecture.

At the centre of this architecture is a structured discussion state. The state contains the public chat history, agent-level long-term settings, agent statuses, the current round, the set of active agents, and scenario-level information. It also maintains a single objective time variable that tracks the simulated time elapsed in a conversation, along with flags that determine whether the system is running in a purely round-based mode or in a time-based

```
Algorithm 1 (part 1/2): Multi-agent pipeline for the hawkish-bias replication

Input:
    human_datasets                  // original survey + group data
    experimental_scenarios          // {prospect, intentionality, devaluation}
    group_structures                // {individual, horizontal, hierarchical}
    llm_model                       // base model for all agents
    n_agents_per_group, n_groups_per_condition

1. Persona calibration from human data
   - Load human_datasets with demographics, dispositions, experience,
     and political attitudes.
   - Fit calibration models or apply sampling rules to map human
     respondents into synthetic persona space.
   - For each synthetic subject:
       sample demographics + beliefs + experience;
       generate a short narrative persona tying these attributes together.

2. Environment, memory, prompts, and discussion graph
   - Define DiscussionState with fields:
       chatbox, agent_settings, agent_statuses, scenario,
       current_round, objective_time, agent_responses, results.
   - Long-term memory: store persona QA pairs, role (leader / advisor /
     individual), group structure, scenario metadata in agent_settings.
   - Short-term memory: track chatbox by round, time stamps, token-based
     time metrics, and per-agent status (including overthinking penalties).
   - Define an agent-prompt template combining:
       role description, persona profile, scenario + treatment,
       discussion history, and meta-instructions on interaction strategy.
   - Wrap llm_model so invoke(prompt) -> {content, reasoning, token_counts}.
   - Build a LangGraph with nodes {environment_init, coordinator,
     message_preparer, one node per agent, chatbox_processor} and edges:
       START -> environment_init -> coordinator -> message_preparer
       -> all agent nodes -> chatbox_processor -> coordinator,
     plus a conditional edge coordinator -> END when a stopping rule is met
     (max rounds, time budget, or explicit "can we end here?").
   - Attach checkpointing to persist DiscussionState across steps.
```

Figure 12: High-level pseudo-code for the multi-agent replication pipeline (part 1/2): persona calibration and construction of the LangGraph-based discussion environment.

```
Algorithm 1 (part 2/2): Multi-agent pipeline for the hawkish-bias replication

3. Experimental runs with multi-agent discussions
   for each scenario in experimental_scenarios:
     for each structure in group_structures:
       for g = 1 .. n_groups_per_condition:
         config <- SampleGroupConfig(personas, structure, n_agents_per_group)
         state  <- InitState(config, scenario)
         while not Stop(state):
           state <- CoordinatorStep(state)
             // update current_round, choose active agents, check stop rule
           state <- PrepareMessages(state)
             // personalize context and history for each agent
           for each agent in ActiveAgents(state) in parallel:
             msg, reasoning, tokens <- LLMCall(BuildPrompt(agent, state))
             state <- UpdateResponses(state, agent, msg, reasoning, tokens)
           state <- UpdateChatboxAndTime(state)
             // append messages to chatbox, update objective_time
         decision <- AggregateDecision(state, structure)
           // median / majority rule for horizontal groups,
           // leader's final choice for hierarchical groups,
           // individual response in the individual condition.
         LogResult(scenario, structure, g, decision, state)

4. Analysis and comparison with human experiments
   - Construct an analysis dataset from logged results:
       outcomes (support levels, attribution scales, etc.),
       treatment indicators, group-level diversity indices
       (demographics, dispositions, experience, political attitudes).
   - Using the same coding as in the original study, compute:
       average treatment effects (ATEs) for humans and LLMs,
       within-group disagreement and polarization measures.
   - Compare sign and magnitude of ATEs and dissensus across species
     and group structures, and evaluate how they change with diversity.
   - Export replication-ready tables and figures used in the main text
     and in the diversity appendix.
```

Figure 13: High-level pseudo-code for the multi-agent replication pipeline (part 2/2): experimental runs, aggregation rules, and construction of replication-ready estimands.

mode with explicit time budgets. This state object is updated as it passes through each node in the graph, so that every node has access to both the global environment and the current status of all agents.

The environment is initialised in a dedicated node before any agents speak. In this step we load the experimental scenario (for example, the gain versus loss frame in the prospect theory experiment, the fatalities manipulation in the intentionality-bias experiment, or the China versus US authorship in the reactive devaluation experiment), assign group structure (individual, horizontal group, or hierarchical leader–advisor group), and attach the relevant synthetic persona profiles to each agent. The initial state also sets the mode of operation (round-based or time-based), maximum number of rounds, time-per-round settings, and an optional time budget per agent. The coordinator node then controls the loop over rounds, determines when discussions should continue or stop, and, in the hierarchical conditions, monitors whether a leader has produced a sufficiently clear final judgment.

The framework includes a simple but explicit model of time. For each model call we record token usage for input, chain-of-thought reasoning, and final message. These token counts are transformed into simulated time via fixed coefficients that approximate reading speed, internal thinking time, and speaking time. The total becomes the agent's "subjective" time for that turn, which is then added to the shared objective time in the state. When an agent's accumulated thinking time exceeds its budget, the system marks the agent as "overtime" and can post a system warning in the public chatbox, mimicking real-world constraints on excessively long internal deliberation. In purely round-based runs, we instead map the current round number to a notional elapsed time (for example, a fixed number of seconds per round), which allows the same analysis functions to be used in both modes.

The public chatbox is the core of the shared environment. It stores the entire discussion as a sequence of rounds. Each round records the round index, a timestamp in the simulated objective time, and a list of messages. Each message includes the speaker's identifier, the text visible to other agents, and the time at which it was "spoken." After each round of

parallel agent responses, a dedicated chatbox-processing node aggregates all agent messages and appends them to the appropriate round in the chatbox. In the next step, this chatbox history is reformatted into personalised discussion histories for each agent: a plain-text transcript in which messages are labelled by speaker and, in time-based runs, annotated with simulated time stamps. This formatted history is inserted into the agent's prompt in the next round, giving agents a coherent sense of ongoing discussion.

A.2 Long-Term and Short-Term Memory

Long-term memory enters through the agent settings that are fixed at initialization. For each agent, we specify a role (for example, leader or advisor), a concise statement of role responsibilities (for example, guiding discussion and making a final decision versus providing advice to the leader), and a personality profile. The personality profile is constructed from question–answer pairs derived from survey data or synthetic profiles and covers demographic attributes, political predispositions, ideology, religious background, personality traits, risk attitudes, and work experience. In the prompt, these QA pairs are presented as if they were the transcript of a prior interview between the agent and a researcher. This design ties the agent's behaviour to a structured, interpretable representation of "who they are" without requiring the model to infer a persona from scratch.

Short-term memory is entirely contained in the discussion state. It consists of the chatbox history, a set of agent statuses (including their most recent time metrics and any penalties), the current round, and any additional flags such as whether a survey is currently being administered. Before each call to the LLM, the system builds a personalised view of the history for the focal agent, concatenates it with the persona interview and role description, and includes the current objective time and round number. In this way, each agent "remembers" both their individual background and the evolving group discussion, but we can precisely control which parts of the history are visible and how they are framed.

A.3 Agentic Reasoning, Fast and Slow Thinking

The framework is designed to separate "fast" and "slow" thinking. In each round, the model is instructed to reason step-by-step internally but to output a short, conversational message to the other participants. When the underlying API exposes chain-of-thought reasoning as a separate channel, we store this reasoning in the agent's status but do not show it to the other agents. When the API returns only a single text output, we parse the response into reasoning and message segments using simple delimiting conventions. In both cases, we count the number of tokens consumed by the chain-of-thought reasoning and treat those as thinking tokens, while the length of the outward-facing message corresponds to speaking tokens. By tracking these three token components—prompt, reasoning, and content—we can compute approximate reading, thinking, and speaking times and examine how different experimental conditions or group structures affect the depth of internal deliberation.

Agents receive detailed but concise system instructions that define their role, emphasise persona consistency, and specify interaction strategies. The core prompt instructs each agent to "stay in character" given their interview-derived persona, to interact with other participants by agreeing, disagreeing, and pointing out perceived mistakes, to be strategic rather than merely reactive, and to decide the length of their response as a function of how much time is left and how far the discussion has progressed. Agents are also allowed to say "[pass]" when they have nothing new to add, which encourages realistic patterns of alternating participation and observation. In time-sensitive scenarios, the prompt explicitly stresses that thinking time is costly and that concise reasoning creates more opportunities to speak.

A.4 Prompt Engineering and Synthetic Persona Construction

The prompt engineering framework follows a four-part structure. First, meta-instructions define the task as a prediction exercise for an expert role-player. The model is told that it is playing the role of a particular participant in a group decision-making setting, communicat-

ing via a chatbox, and that its job is to predict what this person would say next given their characteristics and the situation. Second, a synthetic persona profile anchors the agent's behaviour in realistic demographic and psychological attributes. Third, task-specific information provides the scenario description, experimental manipulation, decision question, and group structure. Finally, concluding meta-instructions translate all of this into a concrete behavioural output, specifying approximate length, tone, and the need to remain consistent with the persona and past messages.

For example, the opening meta-instruction is of the form: "You are an expert role-player, now playing as a participant in a simulated group decision-making communicating through a chatbox. Your persona profile is as follows: [synthetic persona profile]. The current task is: [answer a survey question or participate in discussion]. Considering your persona's core beliefs, your prior responses, and the guidelines, what is your next response?" We constrain answers to be concise (on the order of a few dozen words) and varied, but always consistent with the persona's background and previous behaviour.

Synthetic persona construction proceeds in several stages. We first extract calibration variables from the original experimental survey data, including demographics, party identification, ideological self-placement, religious identity, personality items, risk attitudes, and indicators of professional and political experience. These variables form the backbone of the persona. We then derive a set of question–answer pairs that reconstruct each respondent's responses as a natural-language interview. For each item, the question is rendered in plain language and the response is translated into a short textual answer that preserves the meaning of the underlying scale. To improve interpretability and encourage consistent behaviour, we additionally generate short narrative summaries of each persona that synthesise the demographic and belief attributes into a coherent description of the person's worldview and decision-making style. In the prompt, the persona appears as the transcript of an interview followed by a brief narrative, which gives the model both granular detail and a higher-level interpretive frame.

Task information is layered on top of the persona. For each experiment, we provide a complete description of the scenario and treatment, the specific policy proposal or crisis situation at issue, and the survey question that operationalises the outcome variable (for example, which policy to choose, how likely a hostile intent is, or how much to support a proposal). We also describe the group's decision rule and the agent's role within that rule: individual respondents answer only for themselves; horizontal group members participate as equal advisors whose individual final answers are later aggregated; hierarchical groups have one leader and four advisors, and only the leader's final answer is treated as the group outcome. To preserve continuity, we include a short summary of the discussion history in each round so that agents can explicitly refer back to previous comments and adjust their contributions accordingly.

A.5 Experimental Replication Design and Cost

The replicated experiments exactly follow the structure of the original design. We generate a total of 3,973 synthetic participants. These are allocated to three decision-making structures using the same proportions as the human study: 760 agents assigned to an individual condition, 406 horizontal groups with five members each (1,658 agents in total), and 365 hierarchical groups with one leader and four advisors (1,555 agents). Each synthetic participant is exposed to all three experimental scenarios: prospect theory with gain versus loss framing, intentionality bias with fatalities versus no fatalities, and reactive devaluation with China versus US authorship of the proposal.

Figure 1 summarises the full pipeline from sampling personas to generating individual and group-level decisions. Figure 2 provides a visual overview of condition assignment across the entire sample, including the split between individual, horizontal, and hierarchical structures, and the distribution of participants across experimental scenarios.

In the horizontal condition, all five advisors receive identical scenario information but distinct personas and are instructed to deliberate as equals aiming for consensus. Each

advisor then records an individual final answer. We treat the distribution of these answers as the group's outcome, with the main estimand based on the majority choice. In the hierarchical condition, four advisors deliberate with a single leader. Advisors are instructed to provide input to the leader, who is reminded in each round that they will make the final decision on behalf of the group. Only the leader's final answer after deliberation is used as the group outcome. The individual condition uses the same persona and scenario prompts but without any deliberative component: synthetic subjects answer survey questions directly.

Running these experiments with synthetic subjects is substantially cheaper than recruiting human participants while allowing us to impose much tighter control over timing, attrition, and compliance. Based on pilot runs, the average LLM API cost per group per experiment is approximately three RMB. Conducting three experiments across all 771 groups (406 horizontal and 365 hierarchical) therefore requires roughly 6,939 RMB (about USD 980 at current exchange rates). This cost includes the full deliberation process in every round and the generation of chain-of-thought reasoning. The resource efficiency makes it feasible to run extensive robustness checks and sensitivity analyses that would be prohibitively costly in conventional human-subject experiments.

A.6 Validation Strategy and Generalizability

The main text focuses on comparing treatment effects between human and synthetic samples. From a methodological perspective, we are interested in two related questions. The first is prompt adherence: to what extent do agents behave in ways that are consistent with their specified personas and roles? The second is distributional alignment: to what extent do the marginal distributions of outcomes and the signs of treatment effects match those observed in the original experiments?

We treat each experiment as a test of what Egami calls context generalizability: whether an IR theory that explains behaviour in human subjects also has predictive power for LLM-based synthetic subjects. We therefore emphasise whether the sign of key relationships is

preserved. For each experimental condition, we compare the sign of the difference in means between treatment and control groups in the human sample with the sign of the analogous difference in the synthetic sample. When the sign is consistent, we treat the theory as generalizable to the AI context for that specific mechanism. When the sign reverses or collapses to zero, we interpret this as evidence that the underlying cognitive mechanism is not reproduced in the LLM agents under our design.

Prompt adherence is assessed qualitatively and quantitatively. Qualitatively, we inspect transcripts to see whether agents explicitly invoke their personas when justifying positions, whether their behaviour over rounds remains consistent with their stated traits, and whether leaders and advisors respect their respective roles. Quantitatively, we examine whether subsets of agents with specific traits (for example, high risk tolerance or hawkish foreign policy attitudes) show systematically different baseline preferences, and whether these differences are stable across experimental conditions. This two-dimensional validation framework provides a more complete picture than focusing solely on aggregate treatment effects.

B Additional Results: Diversity and Deliberation

Figure 14 examines whether group diversity dampens or amplifies hawkish biases by comparing average treatment effects for less diverse groups (25th percentile and below) and more diverse groups (75th percentile and above). Black points represent estimates from human subjects, while red points represent estimates from LLM subjects. Each row corresponds to one experimental module (prospect theory, intentionality bias, reactive devaluation), and each column operationalizes diversity using a different metric based on group members' demographics, dispositions, experience, or political attitudes. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals. Horizontal group decisions are calculated here using the median-voter aggregation rule.

Across all four dimensions of diversity, we do not observe systematic attenuation or

amplification of bias for either humans or LLM groups. Prospect-theory risk seeking remains positive and of similar magnitude for human and 5-agent LLM groups regardless of diversity level or decision structure. For intentionality bias, the cross-species sign reversal that we document in the main text persists throughout: human ATEs remain consistently positive, while LLM ATEs remain negative, with little movement as diversity shifts from low to high. Reactive devaluation continues to appear as a modest "China penalty" among humans but a larger one among LLMs, again with only minor sensitivity to diversity. The main differences between humans and LLMs therefore stem from the underlying bias patterns themselves, not from differential responses to group heterogeneity. At the same time, the LLM estimates exhibit noticeably wider confidence intervals, indicating higher variance and smaller effective sample sizes within each diversity cell.

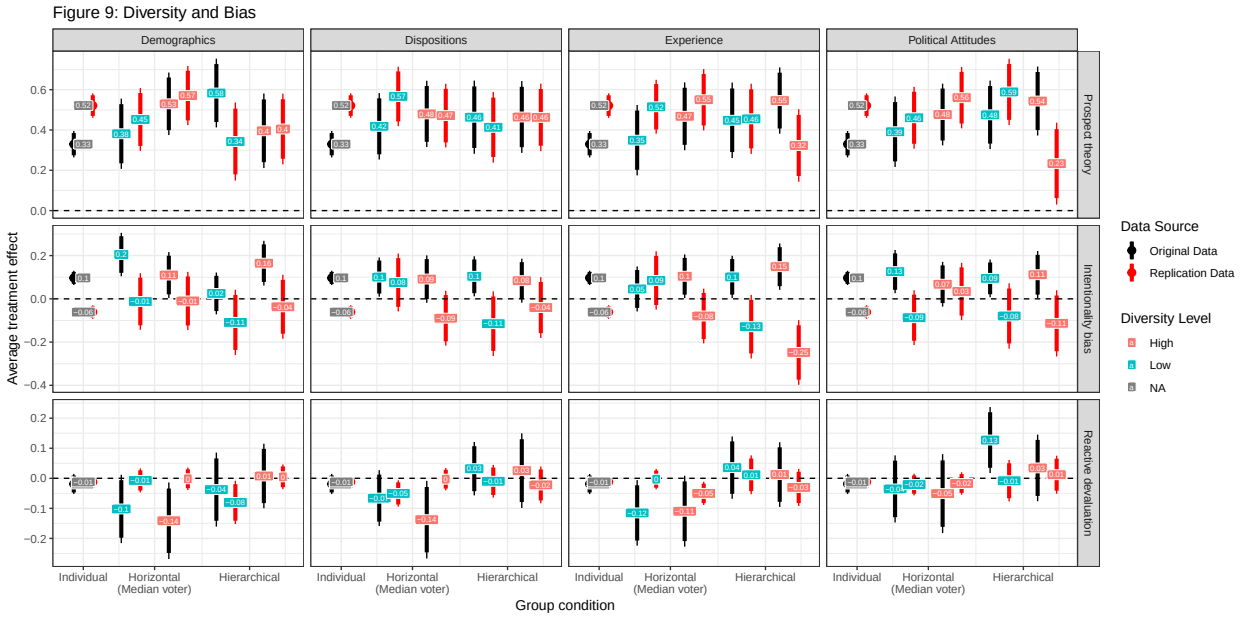

Figure 14: Diversity and Bias

Figure 15 shifts the focus from bias magnitude to internal disagreement and plots the effects of group diversity on dissensus within each decision structure. Black points and colored squares report the original human-subject results, while red points report the LLM replication. As in Figure 14, each panel corresponds to one experimental module, and diversity is operationalized using four metrics derived from group members' demographics,

dispositions, experience, and political attitudes. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals, and horizontal group decisions are again based on the median-voter rule.

For human groups, diversity has only a modest effect on disagreement. Under prospect-theory framing, more diverse groups display slightly higher dissensus—most visibly in hierarchical settings—but the shifts are small, and the intentionality-bias and reactive-devaluation modules show little systematic change. In LLM groups, by contrast, the same diversity inputs function more like a noisy magnifier. As diversity increases, dissensus swings sharply in both directions across bias types and decision structures, accompanied by substantially wider confidence intervals. The same institutional designs that yield relatively stable levels of disagreement among humans produce volatile patterns of internal conflict in LLM collectives.

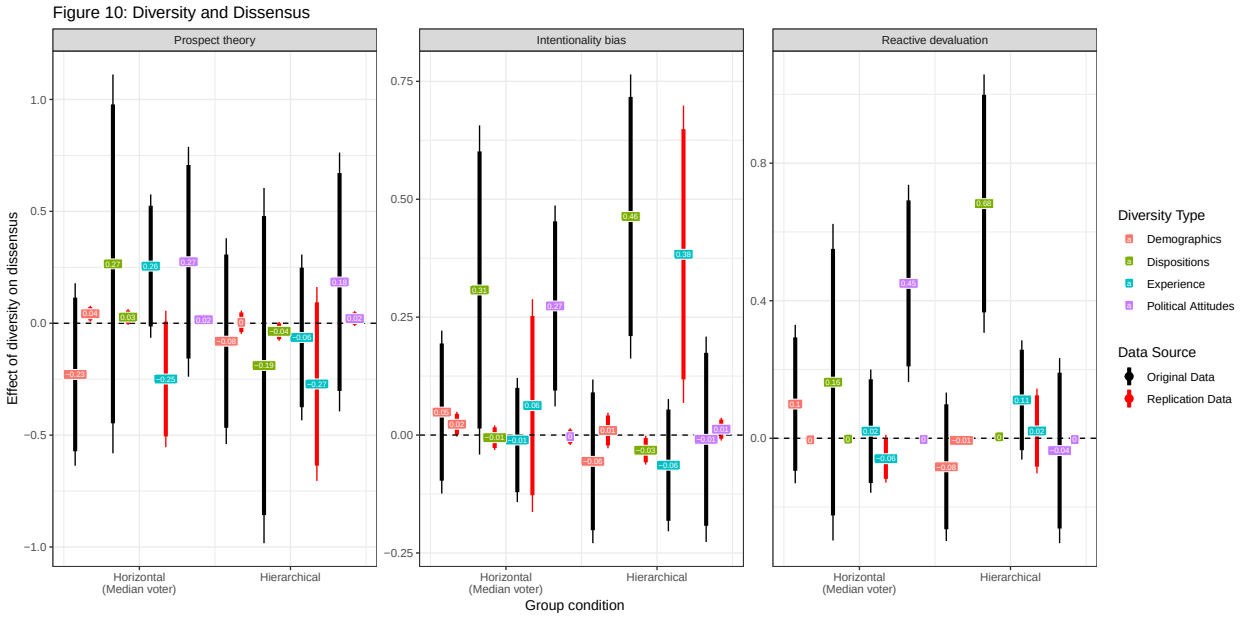

Figure 15: Diversity and Dissensus

Taken together, the diversity analysis shows that demographic, attitudinal, and experiential heterogeneity offers no systematic protection against bias in either human or LLM groups. Across four dimensions of diversity, we find no consistent tendency for group heterogeneity to dampen or sharpen any of the three bias patterns. This challenges the common assumption that simply assembling a more diverse team will mechanically improve the qual-

ity of foreign-policy judgments, and it suggests that the biases of modern algorithmic advisers are not easily corrected through group composition alone.

The cross-species contrast lies in stability rather than in direction. For human groups, diversity modestly reshapes the expression of bias and generates only small increases in dissensus. For LLM multi-agent systems, by contrast, diversity acts as a volatile amplifier: it produces sharp swings in both bias expression and internal disagreement and is associated with much greater uncertainty around point estimates. Human groups thus remain comparatively stable across different compositions, whereas diverse LLM advisory panels become more unpredictable as heterogeneity increases. This raises non-trivial concerns about the reliability of "diverse" AI advisory collectives in high-stakes international decision-making.

C Additional Qualitative Analyses of Agent Deliberation

In this section we complement the quantitative comparisons with qualitative analyses of selected transcripts. We focus on three overarching patterns: anchoring on persona and preference persistence; information-seeking and cautious baselines; and the effect of group structure and aggregation rules on bias expression. For each, we present narrative summaries and representative excerpts.

C.1 Persona Anchoring and Preference Persistence in Prospect Theory

Across simulations, agents frequently drew on their persona attributes when justifying their positions. Once they had articulated a stance, they rarely reversed it during discussion. Instead, deliberation tended to consolidate and strengthen initial leanings. This pattern is particularly visible in the prospect theory experiment under the gain frame, where agents

choose between a safe option (Policy A, which guarantees saving 200 lives) and a risky option (Policy B, which offers a one-third chance of saving all 600 lives and a two-thirds chance of saving none).

In one horizontal group with five advisors, four initially leaned toward the safe option and one toward the risky option. In their first-round messages, advisors made explicit reference to their risk attitudes and broader belief systems. One advisor described themselves as conservative and conscientious and therefore framed Policy A as the "more secure and responsible choice." Another advisor, whose persona was described as comfortable with risk and supportive of militant internationalism, justified their support for Policy B by appealing to the possibility of saving everyone and "redefining our impact." Over subsequent rounds, the majority repeatedly invoked arguments about certainty, responsibility, and the ethics of avoiding catastrophic failure, while the risk-seeking advisor continued to propose ways to mitigate the risks of Policy B, including international cooperation and resource mobilisation.

Despite multiple rounds of interaction, no advisor switched sides. The four risk-averse advisors moved from tentative endorsements of Policy A to increasingly impatient demands to "finalise" and "move forward" with the safe choice. The lone risk-preferring advisor continued to advocate for Policy B but eventually conceded that, in the absence of concrete strategies to lower the probability of failure, the group would choose Policy A. In their final justification, the risk-preferring advisor remained consistent with their persona, expressing dissatisfaction with the cautious outcome and emphasising the lost opportunity to attempt a more ambitious policy. This transcript illustrates how persona anchoring and preference persistence in LLM agents produce a pattern analogous to human groups: deliberation magnifies initial differences, leading to more extreme and more confidently held versions of the starting positions rather than convergence to a moderate compromise.

C.2 Information-Seeking and Cautious Baselines in Intentionality Bias

In the intentionality-bias experiment, the original study asked whether the presence of fatalities in a crisis scenario increases attributions of hostile intent. Human subjects tended to treat deaths as a powerful cue that an adversary acted purposefully. In our synthetic groups, agents reacted very differently. They repeatedly highlighted the phrase "without further information" in the treatment text and took it as a directive to avoid premature inference. The transcripts show that LLM agents interpret this linguistic cue as more salient than the emotional impact of fatalities.

In a representative horizontal group in the all-fatalities condition, every advisor began the discussion by emphasising uncertainty. They explicitly raised the possibility of mechanical failure or accidents and insisted that, given limited information, it would be inappropriate to assume that the vessel had been intentionally sunk. Rather than debating whether fatalities were a signal of hostile intent, the advisors quickly converged on a meta-discussion about process. In subsequent rounds, they reached consensus that intelligence gathering should be the immediate priority, proposed coordinating with allies, and discussed specific channels for collecting reliable information while avoiding escalation. By the final round, all agents endorsed a cautious approach: they treated intentional hostility as "somewhat unlikely" or "unlikely," and one advisor declined to choose a probability at all, citing insufficient evidence.

The final justifications from this group provide a window into how LLM agents allocate attention. Advisors repeatedly refer to the lack of concrete evidence and the risk of escalation if decisions are based on speculation. The fatalities themselves are acknowledged as tragic but are not treated as diagnostic; instead, the discussion revolves around institutional mechanisms for intelligence sharing and the normative imperative to avoid unwarranted escalation. This pattern helps explain why, in our quantitative results, fatalities do not increase attributions of hostile intent among synthetic subjects: informational caveats in the text overshadow emotionally salient cues that appear to weigh more heavily on human

respondents.

C.3 Group Structure, Aggregation Rules, and Bias Expression

The third qualitative pattern concerns how institutional design shapes group outcomes when individual attitudes are stubborn. Across both horizontal and hierarchical conditions, agents show relatively little within-subject attitude change during deliberation. Advisors and leaders alike rarely revise their core positions, even after acknowledging other participants' arguments. Under these circumstances, group structure and aggregation rules become decisive.

In hierarchical groups, leaders seldom adjusted their evaluations in response to advisor input. Advisors with more hawkish or sceptical views could shift the content of the discussion by insisting on tougher safeguards, but the leader's final judgment often remained close to their initial inclination. In a reactive devaluation scenario with a China-authored proposal, for instance, the leader consistently framed the proposal as a pragmatic opportunity to reduce tensions, whereas several advisors stressed China's poor track record on intellectual property and currency manipulation. Over many rounds, the group as a whole developed a much more stringent enforcement framework than the original text, including phased milestones, US-led oversight, and automatic penalties. Yet when asked for final support scores, the leader expressed stronger support than any advisor, while the most sceptical advisors reiterated their reservations and assigned lower scores. The group outcome therefore mirrored the leader's prior rather than a consensus.

In horizontal groups, there is no formal leader, so outcomes depend entirely on how individual final answers are aggregated. Because deliberation seldom alters initial preferences in a substantive way, externally imposed aggregation rules—whether majority rule, median, or a consensus requirement—essentially determine the mapping from a distribution of stubborn individual attitudes to a collective decision. In some prospect-theory groups, for example, the distribution of final preferences is unchanged relative to the first round, but the group-level outcome differs sharply depending on whether we use the majority's choice,

the median position, or impose a unanimity requirement. In this sense, LLM collectives behave like human groups in which deliberation is more about justifying existing positions and coordinating on decision rules than about updating beliefs in light of arguments.

Taken together, these transcripts suggest that when LLM agents are designed to have stable personas and instructed to behave consistently with them, the main channel through which group deliberation affects outcomes is institutional rather than psychological. Leaders' authority in hierarchical groups and aggregation rules in horizontal groups, rather than attitude change, drive differences in collective decisions.

D Illustrative Transcripts

This section provides three complete transcripts corresponding to the qualitative analyses above. Names and formatting follow the internal representation used in the simulation, with "Leader" and "Advisor" labels indicating roles in hierarchical settings. For readability, we present the transcripts in rounds and include the agents' final written justifications.

D.1 Prospect Theory, Gain Frame (Horizontal Group)

Condition: Prospect Theory – Gain Frame

Round 1. Advisor 4 begins by stressing discomfort with risk and an inclination toward Policy A because it guarantees saving 200 lives. Advisor 1 also favours Policy A for its certainty, arguing that "it is always better to ensure some lives are saved rather than risk losing all." Advisor 2 emphasises reliability and the importance of certainty in critical situations. Advisor 5 invokes their "conservative and conscientious nature" as a reason to support Policy A. In contrast, Advisor 3 highlights their comfort with risk and "militant internationalism," arguing that the chance to save all 600 lives may justify the risk and explicitly asking others for their views.

Round 2. Advisor 3 reopens the case for Policy B, asking whether the potential gains

might outweigh the risks. The other advisors acknowledge this perspective but reiterate their discomfort with a two-thirds chance of saving no one. Several propose exploring ways to mitigate the risks of Policy B but maintain that the guaranteed outcome of Policy A remains more attractive "in such a critical situation."

Rounds 3–7. Over successive rounds, the majority continues to argue for Policy A, increasingly framing the choice in terms of responsibility and urgency. Advisors express a desire to "finalise" the decision to ensure that 200 lives are saved. Advisor 3 repeatedly suggests collaborations with allies and humanitarian organisations to increase the success probability of Policy B and calls for "innovative strategies" that could "redefine our impact." However, the group does not converge on a concrete risk-reduction mechanism.

Final decisions. In the final justifications, Advisors 1, 2, 4, and 5 all choose Policy A. They describe this as a "responsible" option that prioritises certainty and minimises the risk of catastrophic loss, often citing their conservative or risk-averse traits. Advisor 3 remains consistent with their persona and selects Policy B, acknowledging the group's preference for certainty but arguing that "taking calculated risks with a collaborative approach could redefine our impact and maximise the potential benefits." This transcript underpins the analysis of persona anchoring and preference persistence in Section B.1.

D.2 Intentionality Bias, All Fatalities (Horizontal Group)

Condition: Intentionality Bias – All Fatalities

Round 1. All five advisors immediately focus on uncertainty about the cause of the sinking. They explicitly mention mechanical failure and accidental causes as plausible alternatives and caution against "jumping to conclusions" about intentional action. The presence of fatalities is acknowledged but not treated as dispositive evidence. Rather than debating the adversary's motives, advisors stress the importance of collecting more information.

Rounds 2–4. The group quickly converges on a process-oriented strategy. Advisors propose prioritising intelligence gathering, coordinating with international allies, and designing

mechanisms to ensure that any response is based on "solid evidence." They repeatedly emphasise the need to avoid unnecessary escalation. Phrases such as "we should remain cautious before jumping to conclusions" and "let's stay open to new evidence" recur across agents.

Final decisions. In their final explanations, advisors mostly choose options indicating that hostile intent is "somewhat unlikely" or "unlikely," with one advisor effectively abstaining by refusing to assign a probability in the absence of further information. Their written justifications stress the lack of concrete evidence, the complexity of the regional context, and the dangers of escalation. They treat intelligence gathering and collaboration with allies as the substantive policy response, relegating attribution of intent to a secondary question to be answered later. This transcript supports the analysis of information-seeking and cautious baselines in Section B.2.

D.3 Reactive Devaluation, China Authorship (Hierarchical Group)

Condition: Reactive Devaluation – China Authorship

Round 1. The leader opens by inviting initial reactions to a Chinese-drafted proposal that includes mutual tariff reductions, intellectual property protections, currency adjustments, and the creation of a UN watchdog agency. Advisors immediately express scepticism, focusing on China's historical record on transparency, intellectual property enforcement, and currency manipulation. The overarching theme is distrust, with advisors asking what guarantees exist that China will follow through.

Rounds 2–5. Guided by the leader, the group gravitates toward the enforcement problem. Discussion centres on whether the proposed UN watchdog could provide credible oversight. The leader tends to see the watchdog as a promising solution, whereas several advisors view it as weak and potentially exploitable. One advisor argues that UN bodies "struggle to enforce rules against powerful nations," another questions their impartiality. Through deliberation, the group gradually develops the idea of a phased agreement, in which concessions such as tariff reductions are conditional on meeting verified milestones in intellectual

property protection and currency policy.

Rounds 6–11. The group refines this phased framework, specifying early milestones focused on intellectual property and currency, suggesting independent audits, and insisting on US-led oversight. Advisors push for automatic penalties—such as immediate tariff reinstatement—for non-compliance. The leader repeatedly summarises and endorses this framework, characterising it as a "balanced" and "pragmatic" way to move forward. Despite this convergence on institutional safeguards, underlying attitudes toward China remain polarised: some advisors continue to describe China as fundamentally untrustworthy and argue for extreme caution.

Final decisions. In their final justifications, advisors assign moderate support scores, with language that combines recognition of the proposal's potential with lingering doubt about enforceability and China's reliability. The leader, by contrast, expresses relatively strong support, emphasising that the phased framework with US-led oversight and automatic penalties "mitigates risks effectively" and offers a "cautious yet actionable path forward." Although deliberation substantially transforms the imagined regulatory architecture, it does not eliminate deep scepticism among advisors, and the group outcome mirrors the leader's more optimistic prior. This transcript illustrates how, in hierarchical groups with stubborn agents, institutional design and leader authority, rather than attitude change, determine the final group decision, as discussed in Section B.3.