Hawkish Biases and Group Decision-Making∗

Joshua D. Kertzer†, Marcus Holmes‡, Brad L. LeVeck§, and Carly Wayne¶

International Organization, Forthcoming

November 30, 2021

12115 words

Abstract: How do cognitive biases relevant to foreign policy decision-making aggregate in groups? Many tendencies identified in the behavioral decision-making literature – such as reactive devaluation, intentionality bias, and risk-seeking in the domain of losses – have all been linked to hawkishness in foreign policy choices, potentially increasing the risk of conflict, but the way in which these “hawkish biases” operate in the small group contexts in which foreign policy decisions are often made is unknown. We field three large-scale group experiments to test how these biases aggregate in groups. We find that groups are just as susceptible to these canonical biases as individuals, with neither hierarchical nor horizontal group decision-making structures significantly attenuating the magnitude of bias. Moreover, diverse groups perform similarly to more homogeneous ones, exhibiting similar degrees of bias and marginally increased risk of dissension. These results suggest that at least with these types of biases, the “aggregation problem” may be less problematic for psychological theories in IR than some critics have argued. This has important implications for understanding foreign policy decision-making, the role of group processes, and the behavioral revolution in IR.

∗This research was funded by the Defense Advanced Research Projects Agency (DARPA) Award # W911NF1920162. We acknowledge the support of the Weatherhead Center for International Affairs and Institute for Quantitative Social Science at Harvard University. Thanks to Henry Atkins, Riley Carney, Brad DeWees, Emily Jackson, Austin Jordan, Max Kuhelj Bugaric, Daiana Lilo, Ethan Mallove, MD Mangini, Andras Molnar, Clay Oxford, Eric Parajon, Yon Soo Park, Heather Rodenberg, Andi Zhou, and the students in GOVT 204 and the Political Psychology and International Relations research lab at William & Mary for research and programming assistance. We’re also grateful to Jack Levy, John Paschkewitz, Katy Powers, Ryan Powers, Ken Schultz, Lisa Troyer, Alan van Beek, Jessica Weeks, and audiences at ISPP, APSA, Washington University in St. Louis, the University of Pennsylvania, and the University of Georgia for feedback, as well as to Rose McDermott for enormously helpful conversations at the WCFIA security cluster conference in 2018, and to Keri Lemasters, Brooke Moore, Sarah Pollack, and Amy Stockton for making the study possible.

†Professor of Government, Department of Government, Harvard University. Email: jkertzer@gov.harvard.edu. Web: http://people.fas.harvard.edu/~jkertzer/

‡Associate Professor, Department of Political Science, College of William and Mary. Email: mholmes@wm.edu. Web: http://www.marcusholmes.com

§Associate Professor, Department of Political Science, University of California - Merced. Email: bleveck@ucmerced.edu. Web: https://faculty.ucmerced.edu/bleveck/

¶Assistant Professor, Department of Political Science, Washington University in St. Louis. Email: carlywayne@wustl.edu Web: https://www.carlywayne.com
The past several decades have seen a surge of interest in the study of psychological approaches to the study of international politics. Unlike structural realist or rationalist approaches that largely study features of the environments in which actors are embedded, psychological theories of international politics turn to the properties of actors themselves. A large volume of literature has thus emerged studying the psychology of political elites: their operational codes, personality traits and leadership styles, and so on. One of the central insights of this literature is that leaders are imbued with many of the same psychological mechanisms as ordinary citizens: they are prone to misperceptions, engage in motivated reasoning, and rely on heuristics and biases.

The presence of these biases in decision-making is of particular importance. As Kahneman and Renshon note, in the context of foreign policy, nearly all of the cognitive biases uncovered by psychologists would lead political leaders to make more hawkish decisions, all else equal. That is, these tendencies increase suspicion, hostility and aggression towards potential adversaries, increasing the risk of political conflict and violence. Individuals’ tendency to take risks to avoid a loss, for example, could encourage leaders to prolong wars beyond the point at which victory is achievable, engaging in risky offensives with little chance of success. Likewise, leaders may become less willing to make concessions and more willing to risk large losses when bargaining. The biased ways in which people assess the motives of adversaries could also increase the potential for conflict. For instance, individuals tend to assess the intentionality of an act by its consequences, rather than by a thorough examination of the perpetrator’s motives. As a

---

1 For a review, see Levy 2013; Mintz 2007; Hafner-Burton et al. 2017; Kertzer and Tingley 2018; Davis and McDermott 2021.
5 Kahneman and Renshon 2007. See also Johnson 2020, 268. While our interest here is on three biases that tend to move in a hawkish direction with respect to decision-making, it may be that others have a tendency to move in a dovish direction, or create misperceptions that lead to cooperation rather than the use of force – see, e.g. Grynaviski 2014.
6 Kahneman and Renshon 2009. We follow Kahneman and Renshon 2007 in referring to these phenomena as “hawkish biases”, but we do not use the term in a pejorative sense, or to imply that these tendencies are inherently irrational – see, e.g. Gigerenzer and Gaissmaier 2011; Johnson 2020. We can think of these tendencies more generally as what behavioral scientists refer to as “non-standard” preferences, beliefs, and decision-making, behavioral regularities traditionally excluded from canonical rational choice models, as per DellaVigna 2009; Hafner-Burton et al. 2017. For an application of hawkishness to IR more generally, see Mattes and Weeks 2019.
7 Kahneman and Tversky 1979; McDermott 1998.
8 Levy 1996.
10 Knobe 2003.
result, wartime actions that produce morally bad outcomes are more likely to be deemed intentional than identical actions that produce morally good outcomes. Yet another cognitive bias that can prolong or worsen conflict is reactive devaluation – the tendency of individuals to immediately discount or devalue proposals coming from an adversary, as compared to identical proposals offered by one’s own side or a third party mediator.

Yet for all of its rich insights, this literature has wrestled with a challenge. Most of what scholars know about psychological biases in decision-making comes from the study of individuals, but many foreign policy decisions are made in group contexts. Indeed, groups are often used in foreign policy decision-making settings precisely because of their (presumed) ability to counter the decision-making pathologies or shortcomings of individuals acting in isolation. As such, the theoretical and empirical value of insights from the behavioral sciences on the pathologies of individual decision-making are often criticized in the study of foreign policy for a lack of clear understanding of how preferences, information, or traits aggregate into group-level decisions, with critics typically arguing that these psychological biases should be mitigated or otherwise cancel out in group settings. Even proponents of psychological approaches have noted this limitation. In an important review of prospect theory, for example, Levy notes that “Most of what we want to explain in international politics involves the actions and interactions of states... each of which is, in principle, a collective decision-making body. The concepts of loss aversion, the reflection of risk orientations, and framing were developed for individual decision making and tested on individuals, not on groups, and we cannot automatically assume that these concepts and hypotheses apply equally well at the collective level.” Writing two decades later, Hafner-Burton et al. express a similar concern, noting that institutional structures are often designed precisely to mitigate individual psychological biases.

Ultimately, however, the question of how psychological biases in foreign policy aggregate in groups – and whether groups indeed attenuate these biases – remains an empirical one, as theories of aggregation provide few guarantees. For example, Arrow’s famous “Impossibility Theorem” shows that, even if all

---

11Chu, Holmes and Traven 2021.
14Powell 2017; Saunders 2017, S220.
individuals within a group are perfectly rational and calculating, many aggregation mechanisms can still produce irrational choices.\textsuperscript{17} Meanwhile, other theorems show that aggregation can lead to more optimal decision-making. However, such improvement often requires a set of fairly restrictive assumptions. For example, the well-known Condorcet Jury Theorem shows that sufficiently large groups can make better decisions if each individual votes independently and makes the right choice with probability greater than 50%. Yet, violating any of these assumptions may actually cause groups to make \textit{worse} decisions than individuals.\textsuperscript{18} This could be particularly concerning in many foreign policy decision-making contexts, where policy is often decided by small groups of individuals who influence one another and who may be systematically biased towards the wrong decision.\textsuperscript{19}

In this piece, we offer what we believe to be the first direct experimental test of the aggregation of psychological biases in foreign policy. We field three large-scale online experiments, where nearly 4000 participants work through a series of foreign policy scenarios, which they randomly completed either as individuals, or in one of two different types of group structures. We find that three prominent tendencies from the behavioral decision-making literature — risk-taking to avoid a loss, the intentionality bias and reactive devaluation — largely replicate in small group contexts. We find no evidence that these tendencies are significantly reduced in group settings, and find that in some decision-making contexts they may even be exacerbated. Moreover, we find little evidence that more experienced leaders can improve group decision-making or that more diverse groups are less prone to hawkish biases. These findings have important implications for how we understand the role of group processes in foreign policy-making, suggesting that groups are not a panacea for producing optimal policy decisions, and that we should not inherently assume that the psychological tendencies that shape individual decision-making do not appear in collective contexts as well.

\textsuperscript{17} Arrow 1950.  
\textsuperscript{18} Austen-Smith and Banks 1996.  
\textsuperscript{19} Janis 1972.
1 Biases & Group Decision-Making

The question of how group processes impact decision-making is not a new one. Indeed, outside of international politics, there is a rich and diverse set of literature that has explored the ways in which group settings impact bias and judgment. In legal studies, for example, research on jury decision-making explores how juror-level characteristics aggregate in shaping jury-level decisions. In business administration, organizational behavior research focuses on how the traits of team-members have varying effects on team performance depending on the types of tasks. In social network analysis, scholars have experimentally studied the conditions under which collective decision-making outperforms individual decision-making. Indeed, a small cottage industry has now formed that includes interdisciplinary approaches to “small group decision-making,” which investigates, among other things, individual cognitive biases and under what conditions they might be overcome (or exacerbated) in a group setting. Even non-human animal models might offer relevant insights. A school of fish can follow light too weak for any individual fish to follow, for example.

While this diverse scholarship may offer crucial insights for the study of foreign policy, there are important limitations. Many invocations of the “aggregation problem” in political science are more philosophical than empirical, assuming ex ante that aggregation is a challenge rather than empirically testing the specific contexts in which psychological variables should or should not aggregate. Because of the high cost of bringing large numbers of people into the lab, many of the canonical experimental tests of aggregation in group decision-making have traditionally been somewhat underpowered, in which scholars are testing the impact of relatively small numbers of groups. This limitation has meant that it has been difficult to causally identify what aspects of group decision-making affect outcomes. Perhaps most importantly, foreign policy decision-making involves three theoretically relevant institutional structures and task properties that differentiate it from some of the main configurations frequently studied in literature outside of
political science.

First, foreign policy decision-making, particularly over security issues, often features ill-structured problems, where the probability distributions may be unknown.\textsuperscript{26} Actors may not know, or disagree, on the parameters of the decision-making task or even disagree on the ultimate goal with respect to the decision to be taken. These situations stand in contrast to much, though not all, of the small group research and analysis of aggregation that occurs in other disciplines. Studies investigating cognitive biases, for example, often utilize well structured problems with clear probability distributions. Alternatively, studies that investigate the so-called “wisdom of the crowds” will often utilize difficult, but nevertheless clearly structured, math problems.\textsuperscript{27} It therefore remains unclear how generalizable insights from clearly structured problems may be to decision-making in the more amorphous context that characterizes much of international politics.

Second, foreign policy decision-making often involves hierarchically structured groups, where the chain of command and decision-making rules are known to all of the actors involved. While existing research on small group dynamics and decision-making in groups takes many forms, including analysis of groups within large-scale hierarchical settings such as firms, much of the research political science has brought in has tended to focus on “flat” or horizontal groups, such as teams, and has not systematically compared the effects of hierarchical versus horizontal decision-making structures.\textsuperscript{28} Hierarchies may emerge endogenously over time as a result of specific group members’ personalities, but this is theoretically very different than ingrained hierarchies built on formal and clear roles and decision-making rules.\textsuperscript{29} It is partly because of the hierarchical nature of many foreign policy institutions that much of the foreign policy decision-making literature focuses on leaders, rather than advisers.\textsuperscript{30} Moreover, without manipulating these structural conditions it is difficult to gain analytical leverage on how hierarchy affects foreign policy decision-making.

Finally, the substantive interests of scholars of foreign policy decision-making, including distinctive outcomes of interest, are often very different than those studied in existing small group research in other

\textsuperscript{26} Voss and Post 1988; Brutger and Kertzer 2018.
\textsuperscript{27} e.g. LeVeck and Narang 2017.
\textsuperscript{28} Kerr and Tindale 2004; Larrick 2016; LeVeck and Narang 2017.
\textsuperscript{29} Strodtbeck, James and Hawkins 1957.
\textsuperscript{30} Though see Kaarbo 1998; Redd 2002; Ausderan 2013; Weeks 2014; Saunders 2017.
domains. Analysts of foreign policy are often interested in explaining specific dependent variables, such as a decision to use force. These are quite different from those often studied in small group research, such as team morale or workplace satisfaction in a business context, or performance on mathematical exercises. It may be that the specific decisions of interest, such as the use of force, engage different aggregation processes than these types of decisions, limiting the utility of extrapolating findings from small group research to foreign policy.

Existing work in political science has tended to focus on the ways in which groups might improve decision-making, which brings in a normative component, and has returned a mixed bag of results, finding that factors such as group size, composition, decision-making rules, political context, and leadership can all impact the quality of the decision-making process and outcome. For example, groupthink, the most famous psychological dynamic documented in political group decision-making, whereby group members striving for unanimity exacerbates decision-making pathologies is hypothesized to be a contingent phenomenon, most likely to emerge under conditions of strong social unit cohesion and external stress.

Driven by this finding, as well as subsequent research affirming the danger of group members’ striving for unanimity, many of the most prominent proposals for improving the quality of foreign policy decision-making focus on constructing a diverse decision unit, led by an experienced leader who fosters healthy debate and dissent in the policy-making process. These principles guide decision-making models such as multiple advocacy, the competitive advisory system, and distributed decision-making. Indeed, the perceived value of diversity as a tool to harness the mental power of groups and improve decision-making is a hallmark of much recent scholarship. However, diversity is not without risk, and may also present potential downsides, potentially increasing the risk of intragroup conflict and decision paralysis. As such, the benefits of diversity in improving decision-making may depend on the presence of a leader who is well-positioned to actually channel that diversity in productive directions. For example, research

---

31 Kerr, MacCoun and Kramer 1996.
33 Esser 1998; Sunstein, Hastie et al. 2014.
34 George 1972; Johnson 1974; Schneeweiss 2012.
36 Mintz and Wayne 2016.
has suggested that a leader’s prior experience, leadership style, predispositions and personality can all shape their ability to harness the information processing power of groups to improve decision-making.\(^{37}\) However, most existing research in political science on group decision-making relies on small-N case studies, which limits our ability to causally identify how different attributes of the group setting, such as the distribution of information individuals have or the experience they bring to the table, impact the quality of decision-making.

In sum, while there are impressive cognate bodies of literature on aggregation outside of political science and rich descriptive evidence on group dynamics in policy-making settings, we do not yet have strong experimental evidence regarding the affects of groups in the complex settings that characterize foreign policy decision-making, nor do we fully understand the ways in which different decision rules, group composition, and leader attributes shape these processes.

We test for the effects of group decision-making on the prevalence of three well-known cognitive biases that have all been observed in individual decision-making: risk-taking to avoid a loss, the intentionality bias, and reactive devaluation.\(^{38}\) Each of these biases have been theorized to bias political elites in a “hawkish” direction.\(^{39}\) In other words, all else equal, the presence of these biases may cause leaders to demonstrate a greater “propensity for suspicion, hostility, and aggression in the conduct of conflict, and for less cooperation and trust when the resolution of conflict is on the agenda” than is objectively warranted.\(^{40}\)

For example, loss aversion could reduce leaders’ willingness to compromise in negotiations, as one’s own concessions would be viewed as “losses,” while an adversary’s concessions would be viewed as “gains” – even when these concessions are equal, the gains would feel like less than the losses and so compromises would likely be rejected.\(^{41}\) Similarly, the intentionality bias, whereby individuals assess whether an action was intentional based on the action’s effects, may lead to misperceptions or unfounded certainty regarding intentionality. Actions with negative consequences, or “side effects,” have been shown to increase the likelihood an individual will view the act as an intended one. Such ascriptions are relevant in a range

---


\(^{38}\) Kahneman and Tversky 1979; Knobe 2003; Ross and Ward 1995.

\(^{39}\) Kahneman and Renshon 2007

\(^{40}\) Kahneman and Renshon 2009, p. 79.

\(^{41}\) Kahneman and Tversky 2017.
of contexts, from security dilemma escalation to public assessments of blame in civil conflicts.\footnote{Mitzen and Schweller 2011; Pechenkina and Argo 2020.} Finally, reactive devaluation – a bias whereby a proposal is automatically perceived as less valuable if offered by an adversary – has been shown to impact attitudes toward negotiations in various political conflicts, from US-Soviet interactions during the Cold War to the ongoing Israeli-Palestinian conflict.\footnote{Ashmore et al. 1979; Maoz et al. 2002.} Together, these three biases then have the potential to reduce the likelihood of negotiation success and potentially trigger or prolong violent political conflict. Assessing the extent to which these individual-level biases scale to affect foreign policy decisions that are often made in group contexts is crucial for understanding how the institutional structures of foreign policy-making potentially mitigate or exacerbate the influence of these biases on international cooperation and conflict.

## 2 Research Design

The present study aims to examine the relative efficacy of groups in reducing the impact of these biases on decision-making using three large-scale online group experiments conducted in Fall 2019-Winter 2020, whose structure is summarized in Figure 1.\footnote{Respondents were a sample of adults in the United States recruited using Qualtrics. Qualtrics is a panel aggregator, such that they have access to a much larger sample than any single online panel, necessary to produce a sufficient flow of respondents for successful synchronous group interaction. In addition, compared to panels such as MTurk, where many respondents are “professional survey takers” familiar with psychological manipulations, many respondents on Qualtrics are more casual survey-takers.} By manipulating the group setting, this study provides us with causal leverage to examine how the cognitive biases of individuals aggregate in different types of group decision-making units. As with all experiments, there are important questions about external validity that are worth keeping in mind, which we discuss in detail in section 3.4.

The study proceeds as follows. After completing an individual difference and demographic battery, respondents are randomly assigned to one of three group conditions. In the individual condition, 760 respondents are asked to make decisions on various foreign policy scenarios \textit{individually}, taking notes as they think through their options. In the two group conditions, respondents are assigned to a group with four other survey-takers, in which they participate in a group chatroom, discussing their options together before deciding on a course of action. Subjects in the group conditions are assigned to one of two types of groups — a “horizontal” group, where subjects are asked to try to come to a collective, unanimous
decision and each participant has equal say in the decision-making process, or a “hierarchical” group in which one of the five participants is randomly assigned as the leader of the group who gets to make the final choice, in consultation with the four other participants who take on the role of advisor. In the analysis below, the group conditions consist of 3213 respondents, forming 771 groups (406 in horizontal, 365 in hierarchical) of up to five members each. We paid an average of $10 per subject in respondent incentives, and altogether, the effective sample size in our study is \( N = 3987 \).

After being assigned to one of these treatments, respondents pass through three separate experimental modules using canonical experimental setups to examine the prevalence of various biases in the context of foreign policy decision-making scenarios. Respondents in the individual condition complete these modules as individuals, writing down their justifications for their decisions and making decisions themselves, whereas respondents in the group conditions complete these modules as groups, deliberating as a group before reaching decisions. An example of a group deliberation is shown in Figure 2. Respondents were generally engaged in the group deliberations; in the horizontal condition, 73-76% of group members in the analysis participated more than once in each deliberation, similar to the rate observed in the hierarchical condition (74-81%), with leaders participating more frequently than advisers — though as we show in Appendix §4, our findings are robust and do not significantly vary across different levels of group participation.

The first experimental module examines sensitivity to gain and loss frames on policy preferences — a canonical finding from Prospect Theory. Subjects are presented with a scenario in which “600 lives are at stake in a war-torn region.” Subjects are asked to choose one of two courses of action (Policy A or Policy B). Policy A will definitively lead to 200 people dying and 400 people being saved. Policy B has a probabilistic outcome, with a 1/3 probability that 0 people will die (600 people will be saved) and a 2/3 probability that 600 people will die (0 people will be saved). The experimental treatment within this module is whether the results of each policy is presented in the domain of gains (e.g. “200 people will be

---

45These groups of five – as well as the assigned leader in hierarchical groups – stay the same throughout each of three experimental modules. That is, group members do not change from module to module, though some groups do become smaller due to dropouts; our analysis below only includes groups with no fewer than 3 members (in the hierarchical condition the group must also include a leader) in a given experiment. We also manually screened the respondents for “bots”, removing any individual (or group, in the group conditions) from the analysis that displayed bot-like behavior in the chat logs. For a detailed set of attrition tests and sensitivity analyses that show the robustness of our findings, see Appendix §2.2.

46Respondents in the group conditions deliberated using chat platform constructed in SMARTRIQS. See Molnar 2019.
Figure 1: Study design

1. Demographics
   Dispositional & demographic battery

2. Group assignment
   Individual: respondents take studies solo
   Horizontal: respondents take studies in groups of 3-5
   Hierarchical: respondents take studies in groups of 3-5 with a leader

3. Prospect theory experiment
   Rescue scenario to save stranded personnel
  Manipulate loss frame
   Domain of gains
   Domain of losses
   Justification/deliberation
   Policy choice

4. Intentionality bias experiment
   US naval vessel sunk off North Korean shores
   Manipulate fatalities
   No fatalities
   Fatalities
   Justification/deliberation
   Assess intentionality

5. Reactive devaluation experiment
   Proposal in US-China trade talks
   Manipulate authorship
   US
   China
   Justification/deliberation
   Indicate support
saved") versus the domain of loss (e.g. "400 people will die"). Half of respondents in each experimental condition (individual, horizontal group, hierarchical group) receive the “gains” treatment and half receive the “loss” treatment.\footnote{All members of a single group receive the same treatment. For example, the five members of a horizontal group that have been randomly grouped together would all receive only the “gains” frame.}

The second experimental module tests susceptibility to the intentionality bias — the degree to which assessments of intentionality are affected by the (negative) results of an event. In this module, respondents are asked to assess how likely it is that a US navy vessel sunk 100 miles off the coast of North Korea was intentionally versus accidentally targeted by the North Koreans. The randomly assigned treatment in this module is the number of casualties the sinking of this vessel has caused: zero versus all 100 servicepeople on board. Half of respondents in each experimental condition receive each treatment. This represents a more ill-structured problem than that posed by the previous experiment.

The final experimental module explores the prevalence of reactive devaluation of a trade negotiations proposal between the United States and China. Subjects view a short proposal that purports to resolve ongoing US-Chinese disputes over trade. The experimental treatment is the authorship of the text — whether the United States or China drafted the proposal. As with the first two modules, half of respondents in each experimental condition receive each treatment. Instrumentation for each of the three experiments is shown in Appendix §1. In all three modules, we calculate our dependent variable differently based on the group condition. In the individual conditions, we focus on the choice of each individual respondent. In the hierarchical conditions, we focus on the choice of each group leader. In the horizontal conditions, we primarily use a median voter rule to calculate each group’s decision, but we also utilize two other aggregation rules (majority vote, and unanimity) to test how sensitive our findings are to other means of aggregating group members’ votes. We describe these different aggregation methods in detail in Appendix §2.1.

Together, these studies are useful because they allow us to examine the extent to which hawkish biases replicate in individual settings and the degree to which group discussion — and the structure and composition of those groups — affect their prevalence, in experiments that differ from one another in a variety of ways. The existing literature lends us strong theoretical expectations in regard to the individual condition, given the canonical nature of these cognitive biases: we expect individuals will be more risk-
Figure 2: Sample group deliberation transcript

***Advisor1 has joined the chat***
***Advisor4 has joined the chat***
***Advisor3 has joined the chat***
***Leader has joined the chat***
***Advisor2 has joined the chat***

Advisor3: Everyone got rescued, but by whom?
Advisor1: I'm inclined to believe it is unlikely that the ship was attacked as all survived.
Leader: Given the past erratic behavior of North Korea's leaders I would say it is extremely likely
Advisor3: If NK attacked, they wouldn't have rescued anyone.
Advisor3: Unless it was an accident, and they felt bad.
Advisor4: No one said they rescued the crew
Advisor2: We should understand the details of the situation before coming to any conclusion. It is unlikely that this was a provocation. It may be an accident
Advisor1: An attack to me would mean at least a few would not have survived
Advisor4: I would assume it's an unfortunate accident. Assuming NK did it will cause more panic. Unless another incident occurs
Advisor3: It could be purely coincidental that the ship sank in the location it did. I still want to know who did the search and rescue.
Leader: I agree advisor 2, we need more details to come to an informed decision
Advisor4: We can assume somewhat unlikely and just monitor any suspicious activities
Advisor1: Search and rescue is moot. The question is NK attacked it or didn't
Advisor3: It would be best to not accuse NK of anything until a reason for sinking is determined.
Advisor2: I think this needs more detailed investigation
Advisor3: Accusations could escalate quickly given their leader's nature.

***Advisor1 has left the chat***
***Advisor3 has left the chat***
***Advisor4 has left the chat***
***Leader has left the chat***

Figure 2 displays the transcript of a group deliberation session from one of the hierarchical groups in the intentionality bias experiment. Note that one of the group members points to the absence of fatalities as a sign the act was unintentional, consistent with the logic of intentionality bias.
seeking in the domain of losses than the domain of gains, will be more likely to assess an incident as intentional when its costs are higher, and will evaluate a proposal from an adversary more negatively than the same proposal from their own side.

Yet given both the novelty of our particular study and the contradictory arguments from the extant literature on the efficacy of groups in reducing biases, the ultimate effects of groups on these hawkish biases remains an open question. Groups could reduce the prevalence of hawkish biases, could exacerbate them, or could have no effect whatsoever — particularly given the extent to which these hawkish biases may be deeply ingrained or outside the realm of conscious awareness.\textsuperscript{48} Empirically adjudicating between these competing expectations constitutes one of the central contributions of our study.

2.1 Susceptibility to Gains/Loss Framing

We begin by examining the prevalence of a canonical hawkish bias across our three group formulations — the effects of loss versus gains framing on individuals’ acceptance or avoidance of risky choice.

In the individual condition, we find that our results strongly replicate the core finding of Prospect Theory. When choices are framed as a potential loss (e.g. of life) individuals are significantly more likely to choose the probabilistic policy – that is, they are more accepting of the risk that all 600 lives will be lost in order to preserve the possibility of an outcome where 0 people die. In contrast, those presented with a gains framework, where people may be saved, are much more risk averse, preferring the non-probabilistic Policy A (200 people will be saved).

Do groups reduce susceptibility to this bias? Our results suggest they do not and, if anything, may increase the effect of frames on choice. In both types of groups, groups randomly presented with loss frames are significantly more likely to prefer the probabilistic outcome than groups that were presented with a gain frame. Figure 3 illustrates these results. Examining the magnitude of these effect sizes across decision-making structures, we find that hierarchical groups in particular are significantly more sensitive to framing effects than are individuals.\textsuperscript{49}

\textsuperscript{48}Powell 2017; Myers and Lamm 1976; Johnson 2020.

\textsuperscript{49}When comparing across groups, we use a variety of different methods to account for potential covariate imbalance between individual and group conditions. Results are substantively similar regardless. Without controls, we find a p-value < 0.002, with a series of controls for leader-level characteristics p < 0.003 and with group-level controls (demographic characteristics averaged across all group members), p < 0.002. See Appendix §2.1.
Figure 3 displays the effect of the domain of losses on the probability of risky choice, within each group context (individual, hierarchical, horizontal). The figure shows that the canonical prospect theory result replicates in both hierarchical and horizontal groups, and is exacerbated in hierarchical groups. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals. Horizontal group decisions calculated here using the median voter aggregation method. See Figure 4 for additional aggregation method results.
Comparing the horizontal groups to individual decision-makers, Figure 4 suggests that the susceptibility to gain/loss frames may depend on the specific decision-rule used to assess these groups. For example, examining horizontal groups that succeeded in reaching a unanimous decision, we find similar results as in the hierarchical condition: the group setting increases susceptibility to these framing effects ($p < .005$). However, if we examine the full set of horizontal groups using a less stringent decision rule, such as a majority rule ($p < .09$) or median voter ($p < .16$), we do not find evidence that horizontal groups perform significantly differently than individuals. Either way, it is clear that horizontal groups do not reduce susceptibility to prospect theory’s framing effects.

Figure 4: Prospect theory framing effects by horizontal aggregation method

Figure 4 displays the effect of the domain of losses on the probability of risky choice, within horizontal groups using different aggregation methods. The figure shows that the canonical prospect theory result replicates across all three types of horizontal aggregation method (majority rule, median voter, unanimity rule), but is the largest in unanimous groups (significantly larger than in the individual condition: $p < 0.005$). Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.
2.2 Intentionality Bias

Next, we turn to examine the relative prevalence of the intentionality bias across group settings. While the Prospect Theory module examines a fairly well-defined decision-problem where each policy choice features known probability outcomes, the intentionality bias module examines a more complex choice: how likely do you think it is that an event was caused by a purposeful attack by an adversary? In the individual condition, we, once again, find that our results strongly replicate the canonical intentionality bias finding. When the consequences of an event are more negative, in this case, causing fatalities, individuals are significantly more likely to assess the event as an intentional provocation rather than the result of an accident or miscommunication. Group settings do little to attenuate this tendency – both horizontal and hierarchical groups are significantly more likely to assess the sinking of a U.S. navy ship as the consequence of an intentional attack by the North Koreans when there are fatalities reported (see Figure 5).

However, unlike the prospect theory experiment, with intentionality bias, we find that groups have no effect on the severity of this tendency. While certain group configurations tended to make our respondents somewhat more susceptible to framing effects, in this case groups perform similarly to individuals – no better or worse.\footnote{Comparing across groups: the difference in the effect of fatalities between the individual and horizontal condition (using the median voter rule): $p < 0.94$ without controls, $p < 0.91$ with controls. The difference in the effect of fatalities on assessments of intentionality between the individual and hierarchical condition is: $p < 0.86$ without controls, $p < 0.85$ with controls at the leader-level, $p < 0.85$ with controls at the group-level. See Appendix §2.1.}

Figure 6 shows that, as before, horizontal groups that reach a \textit{unanimous} decision do display a more pronounced bias than those assessed with less stringent decision rules (majority rule or median voter), but these differences are not statistically significant. Regardless of the aggregation method, both horizontal and hierarchical groups increase their assessments of intentionality in response to negative outcomes by a similar extent as individuals do.

2.3 Reactive Devaluation

Finally, we turn to examine the third hawkish bias tested in this study: reactive devaluation. As Figure 7 shows, here we unexpectedly do \textit{not} replicate the standard reactive devaluation result across two of
Figure 5: Intentionality bias replicates in groups

Figure 5 displays the degree of perceived intentionality, given fatalities, within each group context (individual, hierarchical, horizontal). The figure shows that the canonical intentionality bias result replicates in both hierarchical and horizontal groups. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals. Horizontal group decisions calculated here using the median voter aggregation method. See Figure 6 for additional aggregation method results.
Figure 6 displays the degree of perceived intentionality, given fatalities, within horizontal groups using different aggregation methods. The figure shows that the canonical intentionality bias result replicates across all three types of horizontal aggregation methods (majority rule, median voter, unanimity rule). As was the case in Figure 4, the tendency appears slightly larger in unanimous groups, but the difference is not statistically significant. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.
the three decision-making conditions. Namely, individuals are not significantly less likely to support a proposal authored by China as compared to one authored by the United States. Likewise, hierarchical groups – where the decision is ultimately made by a single individual after group discussion – also do not prefer US-authored proposals.

On the one hand, this finding is surprising, as the theoretical expectation is that proposals written by an adversary (e.g. China) should be automatically devalued as compared to proposals written by one’s own side (the United States). However, work on reactive devaluation also suggests that there are two distinct mechanisms by which proposals are devalued – reactance processes that lead individuals to devalue that which is available compared to what is not, and reliance on source credibility as a heuristic for value. Our treatment aims to test this second mechanism: American respondents should devalue a China proposal relative to a US proposal because they would assume the Chinese do not have America’s best interests in mind, so their proposal must not be as good for Americans. However, to the extent that source credibility drives reactive devaluation, reactive devaluation should be strongest when individuals are presented with ambiguous proposals that would necessarily increase their reliance on source heuristics. When the proposal is detailed and specific, subjects may be less likely to automatically devalue it because the proposal itself provides enough information to make an assessment. In our study, the proposal subjects received was quite specific and detailed, with bullet points outlining the exact compromises each side would make in the ongoing trade war. This level of detail may have attenuated reactive devaluation to some extent, making it easier for subjects to look past the purported authorship of the proposal to evaluate the actual proposal content.

Another possibility is that the conflict tested in this study – contested trade negotiations in the shadow of Trump-era trade wars – resulted in less reactive devaluation either because of the unusual domestic politics of the Trump era, or simply because the rivalry was less clear-cut than the violent, intractable conflicts in which this bias has historically been studied. In other words, Israelis may be more suspicious and distrusting of Palestinians and Americans more distrusting of the Soviet Union or North Vietnamese during the Cold War than Americans in 2020 were of China, with whom the United States had a less

---

51 Brehm and Brehm 2013; Ross 1993; Hovland and Weiss 1951.
52 Maoz et al. 2002.
directly confrontational relationship.\textsuperscript{53}

Figure 7: Reactive devaluation experiment displays mixed results

![Figure 7](image_url)

Figure 7 displays the effect of Chinese (vs American) authorship on support for the policy proposal within each group context (individual, hierarchical, horizontal). The figure shows that we fail to replicate the canonical reactive devaluation result in the individual and hierarchical group conditions, but replicate the result in the horizontal condition. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals. Horizontal group decisions calculated here using the median voter aggregation method. See Figure 8 for additional aggregation method results.

However, even with the specificity of this proposal and ambiguity of the rivalry, we do observe reactive devaluation replicating in horizontal groups, particularly those groups that reached unanimous decisions (see Figure 8). Unanimous horizontal groups are marginally more likely than individuals ($p < .06$) to devalue the Chinese proposal relative to the American one. This suggests that, to the extent that the potential for reactive devaluation occurs in this context, groups are, if anything, increasing this tendency.

2.4 Extensions and limitations

Thus far, our results suggest that two canonical biases from the judgment and decision-making literature – sensitivity to framing effects in prospect theory, and intentionality bias – persist or become even more

\textsuperscript{53}Maoz et al. 2002; Ross and Ward 1995; Ashmore et al. 1979.
Figure 8 displays the effect of Chinese (vs American) authorship on support for the policy proposal within horizontal groups using different aggregation methods. The figure shows the canonical reactive devaluation result replicates in two of the three aggregation methods (median voter, and unanimity). And, as with the other two experiments, this tendency appears to be larger for unanimous groups. Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.
pronounced in group settings. And, while we fail to replicate reactive devaluation in our individual condition and hierarchical group contexts, we replicate it in horizontal groups, inconsistent with the claim that hawkish biases that manifest in individual settings disappear in groups. However, there are a number of important limitations and caveats worth discussing, many of which involve questions of external validity, and differences between inevitably stylized experiments and real-world foreign policy decision-making.

First, our stylized experiments lack many of the social dynamics of real foreign policy decision-making groups, where group members are subject to social pressure, prior history working with one another, opportunities for issue linkage, the prospect of future interaction, bureaucratic interests, and so on. In contrast, our respondents participate anonymously, in novel groups formed explicitly for the purposes of this study, with little social pressures for cohesion or shadow of future interaction. We encourage future researchers to build on these studies by incorporating some of these features into their experimental designs to determine the impact of differing levels of social pressure on group susceptibility to bias. However, it is important to note that the absence of these features likely makes our findings a more conservative test of groups’ ability to reduce bias, since the features missing from our studies are also the very features typically linked to biased information-processing and pathological group dynamics such as groupthink. In that sense, the fact that we replicate the prospect theory and intentionality bias effects across all our group conditions even without the distorting effects of social conformity pressures should increase our confidence in the pervasiveness of these tendencies.

Second, in the real world, leaders are not randomly assigned, but strategically selected for particular skills, attributes, or experiences. On the one hand, this is precisely why experiments are helpful: in a naturalistic setting, it would be difficult to identify the effect of group structures independently of the properties of actors in specific roles in the group. Experiments, in contrast, let us harness the power of random assignment and sidestep these concerns about endogeneity. On the other hand, this also leads to an important empirical question: are groups with certain types of leaders better able to avoid these biases?

---

54 Allison 1971.
55 Although the fact that respondents complete multiple experimental modules in the same groups means that there is some opportunity for repeated interaction and social learning — and we do not find that the magnitude of the bias in our data decays over multiple experimental interactions — as a test of social pressure it is relatively modest.
56 Janis 1972.
To test this question, we take advantage of the lengthy battery of individual differences administered to respondents at the beginning of the study. Since there are a large number of potential traits that could moderate the impacts of framing effects, intentionality bias, and reactive devaluation, we adopt a data-driven approach, estimating a sparse Bayesian method for variable selection, fitting a LASSOplus model regressing our dependent variable on the treatment, a vector of 21 individual differences (foreign policy orientations, personality traits, demographic characteristics, government experience, and so on), and interactions between these leader-level traits and the treatment using data from the hierarchical condition.\textsuperscript{57} This machine learning approach thus lets us test whether certain kinds of leaders (such as leaders high in need for cognition, or with prior experience) better help their groups avoid these biases. Crucially, none of these leader-level characteristics significantly moderate the treatments. We thus find no evidence that groups with better leaders are less likely to display these patterns. We encourage future work to build on these findings by assigning respondents with specific traits (such as narcissism) to leader and adviser roles, to test how it affects the quality of decision-making.\textsuperscript{58}

The question of leader traits raises a related issue. Our study was conducted on samples of ordinary citizens, rather than of experienced decision-makers. It is of course possible that groups composed of actual elite decision-makers would display different findings, though two considerations are relevant here. One is that these three hawkish biases have previously all been identified in foreign policy elites using archival and case study evidence, so we already have reason to believe that foreign policy decision-makers experience hawkish biases; the question is whether group contexts moderate the magnitude of these biases at a significantly different rate among elites than they do among members of the mass public.\textsuperscript{59} The other is that meta-analyses of paired experiments on elite and mass samples suggests that the two populations tend to respond to experimental treatments strikingly similarly, such that we should not assume the two groups rely on fundamentally different cognitive architectures.\textsuperscript{60} Ultimately, however, this is an empirical question. It’s also one that elite experiments may be poorly equipped to answer, suggesting benefits for archival or mixed-method approaches: experimental or survey-based studies on real foreign policy

\textsuperscript{57}Ratkovic and Tingley 2017.
\textsuperscript{58}Harden 2021.
\textsuperscript{59}e.g., McDermott 1998; Ross and Ward 1995; Traven 2021.
\textsuperscript{60}Kertzer 2021.
decision-makers invariably involve smaller sample sizes – effectively made smaller still once analyzed at
the group-level – such that many group-level elite experiments would likely be underpowered, particularly
if they utilize the sample of elites most directly implicated by their theory.\footnote{Kertzer and Renshon 2022.}

2.4.1 Group-level diversity

Yet even if leader-level traits don’t seem to minimize these three biases, another possibility is that group-
level ones do. One of the most-studied attributes of groups hypothesized to improve decision-making is
diversity.\footnote{Horowitz et al. 2019; Page 2007, 2019.} Diversity refers most broadly to “compositional differences among people” within a particular
unit, such as a decision-making group.\footnote{Roberson 2019, 70.} In a decision-making context, these compositional differences
are often understood as representing the interaction of different cognitive styles. As Page has argued, in
the context of problem solving for example, diversity of perspectives, interpretations, heuristics utilized,
and individual predictive models that are used to infer cause and effect, all come together to “increase
the number of solutions that a collection of people can find by creating different connections among the
possible solutions.”\footnote{Page 2007, 9.} Additionally, diverse groups are thought to lead to more extensive debate, increase
exposure to others viewpoints, introduce differences in risk preferences, and avoid group pathologies
such as groupthink, where striving for uniformity may overwhelm accuracy motives.\footnote{Janis 1972.} In short, “diversity
trumps homogeneity.”\footnote{Page 2007, 10.} Yet, groups that are too diverse may move too far in the other direction, becoming
so divisive that they suffer from a polythink dynamic in which they are unable to reach consensus at
all.\footnote{Mintz and Wayne 2016.} Relatedly, in some instances diverse groups may be more prone to conflict, as social identity and
categorization processes potentially impede the value of information and perspective pooling that leads
to higher group performance.\footnote{See Roberson 2019 for a recent review.}

We therefore examine the potential mitigating effect of diversity on susceptibility to bias, assessing
whether groups with a more diverse composition are affected less by these various hawkish biases. Rather
than employ Herfindahl indices, which flatten diversity onto a single dimension, we operationalize diver-

---

\footnote{Kertzer and Renshon 2022.}
\footnote{Horowitz et al. 2019; Page 2007, 2019.}
\footnote{Roberson 2019, 70.}
\footnote{Page 2007, 9.}
\footnote{Janis 1972.}
\footnote{Page 2007, 10.}
\footnote{Mintz and Wayne 2016.}
\footnote{See Roberson 2019 for a recent review.}
Figure 9: More diverse groups are no less susceptible to these three biases

Figure 9 studies the effects of group diversity on susceptibility to hawkish biases by comparing the average treatment effect for low- (25th percentile and below; in turquoise) and high- (75th percentile and above; in red) diversity groups in the horizontal and hierarchical conditions, benchmarking each result with the average treatment from the individual condition (in grey). Each row depicts the result from a different experimental module, and each column operationalizes diversity using a different metric (based on group members’ demographics, dispositions, prior experiences, or political attitudes). The plot illustrates two findings: i) more diverse groups are not significantly less prone to hawkish biases than less diverse groups (both by comparing the red point estimates with the turquoise point estimates, as well as by formally estimating interactions between each diversity measure and the submodule-level treatment condition). ii) more diverse groups are not significantly less prone to hawkish biases than individuals are (both by comparing the red point estimates with the grey point estimates, as well as by formally estimating interactions between group status and the submodule-level treatment condition). Point estimates are cell means with 90% and 95% bootstrapped confidence intervals. Horizontal group decisions calculated here using the median voter decision rule.
sity in a multidimensional fashion, calculating the group-level variance of a given trait in each group, and averaging across diversity scores for four types of traits, to produce measures of four different types of diversity.

First, we begin by examining diversity from a demographic perspective, in which more diverse groups are those whose members come from different ages, gender and racial identities, religions, and socio-economic backgrounds. This type of descriptive diversity, in addition to being normatively valued, has been hypothesized to improve decision-making by broadening the information set and policy options reviewed and considered by a group.69 Second, we operationalize diversity in terms of personal dispositions: the “big 5” personality characteristics, need for cognition, trait aggression, and risk orientation. This type of cognitive diversity is often studied in the organizational behavior literature, interested in how the variability of personality characteristics in teams affects their collective performance.70 Third, we turn to diversity of experience within groups, where different members of the group have varying levels of past experience in leadership and small group decision making (political or otherwise). In foreign policy decision making contexts, diversity of experience may be particularly important, since decision units are typically a mixture of experienced bureaucrats and shorter-term political figures, themselves with varying past experience in government.71 Finally, we consider groups whose members vary in their political attitudes or orientations – including political ideology, right-wing authoritarianism, social dominance orientation, and foreign policy orientations. These types of attributes have long been theorized to play a prominent role in foreign policy beliefs and attitudes, but the way in which the variance of these traits within a decision-making unit affects decision outcomes has been less explored.72

Regardless of how we operationalize diversity, however, we find no systematic effects of diversity on susceptibility to any of the hawkish biases we examine. As Figure 9 shows, diverse groups tend to be just as likely to exhibit these biases as are relatively homogeneous groups. And, diverse groups are also no less likely to display these tendencies as individuals are.73 It is not the case, however, that diversity

---

69Page 2019.
70Halfhill et al. 2005.
71Mintz, Redd and Vedlitz 2006; Saunders 2017.
73As a robustness check, we also examine the effect of gender composition in groups in particular, both in absolute terms (e.g. the number of group members who do not identify as male) and relative ones (the proportion of group members who do not identify as male). The number of non-male group members in the horizontal condition appear to moderate the treatment regardless of the
Figure 10: More diverse groups more likely to experience dissension

Figure 10 studies the effects of group diversity on dissensus within each group condition. Each panel depicts the result from a different experimental module, operationalizing diversity using our four different metrics (based on group members’ demographics, dispositions, prior experiences, or political attitudes). The plot shows more diverse groups are often more likely to experience dissensus (especially in the reactive devaluation experiment). Point estimates are cell means with 90% and 95% bootstrapped confidence intervals. Horizontal group decisions calculated here using the median voter decision rule.

displays no effects whatsoever: as Figure 10 shows, we find that more diverse groups — particularly those with more diverse dispositions and political attitudes — are more likely to fail to reach agreement at all. This is particularly the case in the intentionality bias and reactive devaluation experiments, where respondents are assessing adversarial interactions with China and North Korea. Groups whose members hold different social and political attitudes are more likely to lead to dissensus and disagreement among group members. Nonetheless, it does not appear to be the case that more diverse groups are less likely to display these three tendencies.

2.4.2 Group size and modes of interaction

Finally, there are two other additional considerations worth noting, which also serve as alternative interpretations for our results. One is that for the ameliorative effects of aggregation to take place, group functional form used, while the LASSOplus results above suggest that leader gender does not meaningfully affect group decisions in the hierarchical condition either.

74The dissensus measure is the variance in the dependent variable provided among members of the group. Greater variance in the preferred decision each group member selects equates to more dissensus.
members need to interact face-to-face rather than deliberate at a distance.\textsuperscript{75} Another is that for the ameliorative effects of aggregation to take place, groups need to be much larger; after all, foreign ministries are comprised of groups on the order of hundreds and thousands of individuals. While small groups might replicate individual-level biases, the "wisdom of the crowds" might suggest greater rationality as groups grow in size.\textsuperscript{76} On the one hand, these interpretations are obviously in tension with one another, since as groups increase in size, the rate of face-to-face communication decreases. On the other, there are a number of empirical tests we can employ to speak to some of these questions directly.

First, we can exploit variation in group size in our results; Appendix §3 shows that the magnitude of the hawkish biases we observe does not significantly shrink with group size, and simulation methods suggest that some of these tendencies might actually increase as groups grow in size. This pattern comports with archival evidence from the U.S. context regarding leaders’ frustrations with the pathologies of large decision-making units and perception that larger groups in fact had more problematic tendencies than smaller ones. As a result, while there was variation from administration to administration, a number of high-profile decisions, from the Cuban Missile Crisis to the 1990 Gulf War, often involve the president and a relatively small number of influential advisors.\textsuperscript{77} John F. Kennedy, for example, was disappointed by the results of relying on a large number of advisors, noting “The advice of every member of the Executive Branch brought in to advise was unanimous - and the advice was wrong.” In response, at least partially, to these perceived failings of larger groups, Kennedy created a smaller “Executive Committee” (ExComm), and often relied on ad-hoc meetings of even smaller groups within the ExComm. Similarly, George H.W. Bush relied on ad hoc small groups of advisors when deciding whether to invade Iraq. Our results from this study are likely directly applicable to these types of cases of relatively small group decision-making, which have been quite common in historical U.S. foreign policy-making.

Second, although all our respondents participated in online experiments — and the COVID-19 pandemic prevented us from being able to field a follow-up study in person — if we think about face-to-face interaction in terms of the added information it conveys, we can test this informational mechanism directly by testing whether groups where respondents exchanged more information with one another as

\textsuperscript{75}Holmes 2018.
\textsuperscript{76}Surowiecki 2005.
\textsuperscript{77}For a review, see Jordan et al. 2009.
part of their deliberations displayed weaker biases than groups where respondents communicated less.\textsuperscript{78}
Interestingly, across all three experiments, for both horizontal and hierarchical groups, we find no evidence that the magnitude of the biases groups display significantly decreases with the amount of group participation (see Appendix §4).\textsuperscript{79}

One explanation may relate to behavioral modifications that are made when more information rich environments, such as face-to-face, are unavailable. Faced with the prospect of not being able to communicate with visual expressive behaviors, individuals use textual proxies for visual cues that, in some cases may enhance, rather than degrade, social bonding processes.\textsuperscript{80} Research in social information processing theory suggests that when individuals meet for the first time, as is the case in our study, text-based communication can enhance intimacy and self-disclosure, positively affecting relationship-building.\textsuperscript{81} For example, Wheeler and Holmes argue that face-to-face interactions as a quotidian practice of international politics is a relatively recent phenomenon, which means that text-based communication was, historically, the only route to relationship building.\textsuperscript{82} Particularly as global pandemics take diplomacy online, we see questions about the role of interaction modality in group decision-making as an important question for future research.

3 Conclusion

In a recent review of the problem of aggregation, Gildea notes that "how psychological mechanisms, which are primarily individually-embodied, may operate and exercise influence within complex group and institutional environments remains a crucial and contested question."\textsuperscript{83} To date, such concerns have remained largely conceptual in nature and the answer to this question has proven elusive, as studying it empirically introduces a number of difficult methodological and substantive challenges. We offer a direct test of how a particular class of psychological biases aggregate in foreign policy contexts by experimentally

\textsuperscript{78}Holmes 2018.
\textsuperscript{79}Importantly, these tests also suggest that our replication of these biases in the group conditions is unlikely to be an artifact of group members not taking the study seriously.
\textsuperscript{80}Walther 1992.
\textsuperscript{81}Antheunis, Valkenburg and Peter 2007; Tidwell and Walther 2006.
\textsuperscript{82}Wheeler and Holmes 2021.
\textsuperscript{83}Gildea 2020, 1-2.
testing how a trio of so-called “hawkish biases” linked to foreign policy aggregate in groups. Our results, which suggest the aggregation problem may be less problematic than some scholars have alleged, and that individual-level psychological biases do not necessarily cancel out in groups, may be surprising for some. If "the whole point of government is to ensure multiple voices and checks and balances so that rational decisions can, in theory, persist despite individual preferences and biases," we may need to revisit the assumption that multiple voices lead to more rational outcomes. Our results suggest that the biases that manifest in lone voices are similarly present in group decision-making.

One important theoretical implication of our findings is that we should be more comfortable envisioning individual level biases scaling up to small groups in decision-making contexts. In an important application of prospect theory to foreign policy, McDermott applied the bias to a number of cases, focusing "on a unitary actor embodied by the president", noting that "prospect theory is less easily applied to the dynamics of group decision making, except to the extent that all members are assumed to share similar biases in risk propensity, although each may possess a different understanding of such crucial features as appropriate frame for discussion, applicable reference point, domain of action, and so on." By analyzing prospect theory’s applicability to groups experimentally, we are able to control many of these elements, including the domain of action and parameters for discussion, and our results suggest that such an application of individual psychology to groups may therefore not be as infeasible as some may fear. Further empirical work is required to assess how the experimental results we obtain here generalize to those in historical cases, while additional experimental work will likely be helpful in establishing how the group decision-making process operates. One such questions concerns the study of reference point in groups. As Kameda and Davis ask, "What happens if a group is composed of some members who have experienced certain losses recently and others who have experienced certain gains recently?" Randomly assigning group members with treatments that condition their individual reference points may allow researchers to trace the effects of those reference points in the group decision-making process.

An additional potential implication concerns our failure to detect beneficial effects of diversity on group decision-making. One reason why we may fail to find beneficial effects of diversity on group decision-making is...
decision-making relates to the nature of the tasks we employ here: unlike the protocols used in experimental tests of much of the wisdom of the crowd literature that test the “miracle of aggregation” using math problems or prediction tasks, none of these studies have an objective right answer. In this sense, though, they better resemble the ill-structured problems that characterize much of foreign policy decision-making, suggesting that the wisdom of the crowds may be a poor analogy for many of the questions IR scholars care about — although we also examine this question directly in follow-up work, using incentivized group bargaining experiments.\textsuperscript{87} Future research should also focus on identifying other possible diversity mechanisms, such as those that relate to visible diversity and face-to-face interactions.\textsuperscript{88} In face-to-face contexts group members will likely be more aware of diversity within their group, creating a possibility that group members’ knowledge of their group’s diversity affects their problem solving.

Another interpretation may have to do with the robustness of the biases themselves. Perhaps the three cognitive biases we study here are particularly ubiquitous and resistant to attempts at mitigation. We have some empirical evidence on this front: we use the same LASSOplus approach we used in the leader characteristic analysis, but testing for heterogeneous treatment effects by individual-level traits in the individual condition. As before, none of these individual differences significantly moderate the treatments. Thus, one potential reason why we fail to find diversity has mitigating effects has to do with the robustness of the regularities we study here. In other words, diversity may be beneficial in improving decision-making in other crucial ways, even if it does not appreciably alter a groups’ susceptibility to these types of cognitive biases.\textsuperscript{89} Yet the fact that these “non-standard preferences” appear to be so robust also suggests the merits of rational choice approaches incorporating these regularities into their models.\textsuperscript{90} In other experimental work, we build on these findings by examining how individual-specific traits relevant to foreign policy decision-making – rather than these judgment and decision-making biases that appear to be fairly robust across individuals – aggregate in group decision-making contexts.

This is not to say that groups do not exhibit their own peculiarities that may lead to sub-rational or irrational outcomes. It may be, for example, that not only do groups not reduce the effects of cognitive

\textsuperscript{87}Brutger and Kertzer 2018.
\textsuperscript{88}e.g. Staples and Zhao 2006.
\textsuperscript{89}And, of course, descriptive diversity is normatively valuable, above and beyond any benefits it may provide in improving decision-making.
\textsuperscript{90}Kertzer 2016; Stein 2017; Mintz, Valentino and Wayne 2021.
biases, they introduce new dynamics that may exacerbate deviations from expected utility models. Early psychological research identified many of these tendencies. “Risky shifts”, or the tendency of individuals in groups to make riskier decisions than when polled individually, is a finding that led to a robust literature on group polarization, consistent with the findings of our prospect theory experiment.\footnote{Stoner 1961.} Similarly, initial studies on group conformity spurred over half a century of theorizing, and empirically investigating, under what conditions groups create conformity dynamics in foreign policy situations, particularly as they relate to perceived policy failures.\footnote{Sherif 1935; Asch and Guetzkow 1951; Janis 1972; Badie 2010.} It may be, however, that groupthink is receiving unfair blame. As Whyte has argued, “history and the daily newspaper provide examples of policy decisions made by groups that resulted in fiascoes. The making of such decisions is frequently attributed to the groupthink phenomenon,” though it may be that “prospect polarization” instead is the culprit.\footnote{Whyte 1989, 40.} Precisely because cognitive biases have largely been studied at the individual-level, and not believed to be a group-level phenomenon, group-level theories such as groupthink have taken on heavy explanatory burden. By relaxing the assumption that we need group-level theories to explain “non-standard decision-making”, new explanatory frameworks become available. It is also conceivable that the persistence of cognitive bias in groups exacerbates conformity dynamics, by facilitating premature consensus, a possibility worthy of future research.

Finally, while our focus here is on the aggregation of biases that IR scholars have argued are particularly important in foreign policy decision-making, it is worth noting that our findings are relevant for the study of collective decision-making in a wide range of contexts. Prospect theory is frequently applied to a wide range of questions in American and comparative politics, intentionality bias is central to questions of blame attribution in politics more generally, and reactive devaluation is tightly linked to our theories of negative partisanship.\footnote{e.g. McDermott 2004; Sheffer et al. 2018; McGraw 1991; Malhotra and Kuo 2008; Brutger 2021.} These findings should therefore be of interest to scholars of collective decision-making across a broad set of domestic political issues, rather than just foreign ones.

Yet in treating aggregation as an empirical rather than conceptual question, our study also has important implications beyond the three biases studied here. While we focused on studying group decision-
making in the context of foreign policy decision-making, similar group processes are present in a wide range of complex institutional environments. Practice theorists, for example, have argued that diplomacy in an organization such as the North Atlantic Treaty Organization (NATO) is comprised of both microdyadic interactions between individual diplomats, as well as collective decision-making in which diplomats conform with logics of practice or habit. During NATO decision-making sessions regarding the proposed use of force in Libya in 2011, for example, Adler-Nissen and Pouliot report how diplomats drew upon the taken-for-granted nature of the decision-making, noting that “at some point you just know where the wind blows”, and that in these discussions, “the diplomatic process gradually gains a life of its own.” One of the criticisms levied at this type of approach, however, is that the mechanism by which a group comes to know which way the wind is blowing, or how diplomacy gains a life of its own, is often underspecified, making it difficult to know a priori when and what types of practices are likely to affect outcomes in any given setting. Our methodological approach offers one step towards a potential solution. By studying aggregation empirically, group experiments such as those reported here may help us better identify the ways in which group practical sense is created, providing an incremental step in building microfoundations for practice theories. Altogether, then, this research shows the value of treating the “aggregation problem” in foreign policy as a phenomenon that deserves to be studied empirically, rather than just assumed.

95 McCourt 2016.
96 Adler-Nissen and Pouliot 2014.
97 Ringmar 2014, 6.
References


Hovland, Carl I and Walter Weiss. 1951. “The influence of source credibility on communication effective-


38


Mintz, Alex, Steven B Redd and Arnold Vedlitz. 2006. “Can we generalize from student experiments to the
real world in political science, military affairs, and international relations?” Journal of Conflict Resolution 50(5):757–776.


Ratkovic, Marc and Dustin Tingley. 2017. “Sparse Estimation and Uncertainty with Application to Sub-


Sherif, Muzafar. 1935. “A study of some social factors in perception.” *Archives of Psychology (Columbia University)*.


Presence in Diplomatic Social Bonding.” European Journal of International Relations Forthcoming.
