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 hypothesized to increase hawkishness in foreign policy preferences, 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 (n=3973) to test how these tendencies translate in both hierarchical and horizontally-structured groups. We find that groups are just as susceptible to these hawkish biases as individuals, with neither group decision-making structure 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 and decision paralysis. These results suggest that the “aggregation problem” may be less problematic for psychological theories in IR than some critics have argued, as many core cognitive biases known to affect individual decision-making appear to have a similar impact in small group contexts. This has important implications for understanding the psychology of political conflict, foreign policy-making, and the role of group processes in international politics.
The past several decades have seen a surge of interest in the study of psychological approaches to the study of international politics (for a review, see Levy, 2013; Mintz, 2007; Hafner-Burton et al., 2017; Kertzer and Tingley, 2018; Davis and McDermott, 2020). Unlike structural realist or rationalist approaches that largely study features of the environments in which actors are embedded (Waltz, 1979; Lake and Powell, 1999), psychological theories of international politics turn to the properties of actors themselves (Larson, 1985; McDermott, 1998; Vertzberger, 1998; Saunders, 2011; Rathbun, 2014; Kertzer, 2016; Renshon, 2017; Yarhi-Milo, 2018; Holmes, 2018; Landau-Wells, 2018; Powers, 2021). A large volume of literature has thus emerged studying the psychology of political elites: their operational codes (Leites, 1951; George, 1969), personality traits and leadership styles (Greenstein, 1969; Etheredge, 1978; Hermann, 1980), 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 (LeVeck et al., 2014; Sheffer et al., 2018; Kertzer, Forthcoming): they are prone to misperceptions (Jervis, 1976; Stein, 1988), engage in motivated reasoning (Baekgaard et al., 2017; Kertzer, Rathbun and Rathbun, 2020), and rely on heuristics and biases (Poulsen and Aisbett, 2013; Brooks, Cunha and Mosley, 2015; Johnson, 2020).

The presence of these biases in decision-making is of particular importance. As Kahneman and Renshon (2007) 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.1 That is, these tendencies increase suspicion, hostility and aggression towards potential adversaries (Kahneman and Renshon, 2009), increasing the risk of political conflict and violence.2 Individuals’ tendency to take risks to avoid a loss (Kahneman and Tversky, 1979), 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 (McDermott, 1998). Likewise, leaders may become less willing to make concessions and more willing to risk large losses when bargaining (Levy, 1996). The biased ways in which people assess the motives of adversaries could also increase the potential for conflict (Jervis, 1976). For instance, individuals tend to assess the intention-

---

1See 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.

2We 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 (DellaVigna, 2009; Hafner-Burton et al., 2017). For an application of hawkishness to IR more generally, see Mattes and Weeks (2019).
ality of an act by its consequences, rather than by a thorough examination of the perpetrator’s motives (Knobe, 2003). As a result, wartime actions that produce morally bad outcomes are more likely to be deemed intentional than identical actions that produce morally good outcomes (Chu, Holmes and Traven, Forthcoming). 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 (Ashmore et al., 1979; Maoz et al., 2002).

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, often, groups are used in foreign policy decision-making settings precisely due to their (presumed) ability to counter the decision-making pathologies or shortcomings of individuals acting in isolation (‘t Hart, Stern and Sundelius, 1997). 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 even traits, aggregate into group-level decisions (Powell, 2017), typically arguing that these psychological biases should be mitigated or otherwise cancel out in group settings (Saunders, 2017, S220). Even proponents of psychological approaches have noted this limitation. In an important review of prospect theory, for example, Levy (1997, 102), 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. (2017, S18-S21) express a similar concern, noting that institutional structures are often designed precisely to mitigate 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 individuals within a group are perfectly rational and calculating, many aggregation mechanisms can still
produce irrational choices (Arrow, 1950). 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% (Austen-Smith and Banks, 1996). Yet, violating any of these assumptions may actually cause groups to make worse decisions than individuals (Austen-Smith and Banks, 1996). 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 (e.g. Janis, 1972).

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 — 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.

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 (Devine et al., 2001). 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 (Moynihan and Peterson, 2001). In social network analysis, scholars have experimentally studied the conditions under which collective decision-making outperforms individual decision-making (Bernstein, Shore and Lazer, 2018). 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 (Berdahl et al., 2013), 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 (e.g. Mercer, 1995, 237-238; Wendt, 2004; Johnson, 2015; Powell, 2017; Gildea, Forthcoming). 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 (e.g. Lewin, Lippitt and White, 1939). 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 (Voss and Post, 1988; Brutger and Kertzer, 2018), where the probability distributions may be unknown. 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” (e.g. LeVeck and Narang, 2017) will often utilize difficult, but nevertheless clearly structured, math problems. 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. Much of the existing research on small group dynamics and aggregation of preferences in groups has tended to focus on “flat” or horizontal groups, such as teams. In these cases, the institutional structure and chain of command may not be provided. Hierarchies may emerge over time as a result of specific group members’ personalities (Strodtbeck, James and Hawkins, 1957), but this is theoretically very different than ingrained hierarchies built on formal and clear roles and decision-making rules. 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 (though see Kaarbo, 1998; Redd, 2002; Ausderan, 2013; Weeks, 2014; Saunders, 2017).

Finally, the substantive interests of scholars of foreign policy decision-making, including distinct traits, group composition characteristics, and outcomes of interest, are often very different than those studied in existing small group research in other domains. Analysts of foreign policy are often interested in explaining specific dependent variables, such as a decision to use force, sets of traits that may have an effect in crafting those decisions, such as hawkishness or military experience, and the ways in which information is distributed across group members. These are quite different than those often studied in small group research, such as team morale or workplace satisfaction in a business context, or performance on mathematical exercises. Similarly, 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.

Indeed, 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 (Kerr, MacCoun and Kramer,
For example, groupthink (Janis, 1972), 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 (’t Hart, Stern and Sundelius, 1997), most likely to emerge under conditions of strong social unit cohesion and external stress.

Driven by this finding, as well as more recent research affirming the danger of group members’ striving for unanimity (Esser, 1998; Sunstein, Hastie et al., 2014), 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 (George, 1972), the competitive advisory system (Johnson, 1974), and distributed decision-making (Schneeweiss, 2012). Indeed, the perceived value of diversity as a tool to harness the mental power of groups and improve decision-making (Horowitz et al., 2019; Page, 2019) 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 (Mintz and Wayne, 2016). As such, the benefits of diversity in improving decision-making may depend on the presence of a leader that is well-positioned to actually channel that diversity in productive directions. For example, research has suggested that a leader’s prior experience (Horowitz and Fuhrmann, 2018; Saunders, 2017), leadership style (Hermann and Preston, 1994; Preston, 2001), predispositions and personality (Schafer and Crichlow, 2010) can all shape their ability to harness the information processing power of groups to improve decision-making. 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.
2 Research Design

The present study thus aims to examine the relative efficacy of groups in reducing the impact of “hawkish biases” on decision-making using three large-scale online group experiments conducted in Fall 2019-Winter 2020, whose structure is summarized in Figure 1. By manipulating both the group setting and decision-rule, this study provides us with causal leverage to examine how the cognitive biases of individuals aggregate in different types of group decision-making units.

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 individually, taking notes as they think through their options. In the two group conditions, respondents are assigned to a group of 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 mod-

---

3 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.

4 These 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.
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 leader

3. Prospect theory experiment

   Rescue scenario to save stranded personnel
   Manipulate loss frame
   Domain of gains
   Justification/deliberation → Policy choice
   Domain of losses

4. Intentionality bias experiment

   US naval vessel sunk off North Korean shores
   Manipulate fatalities
   No fatalities → Justification/deliberation → Assess intentionality
   Fatalities

5. Reactive devaluation experiment

   Proposal in US-China trade talks
   Manipulate authorship
   US → Justification/deliberation → Indicate support
   China
ules 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.\(^5\) A sample group deliberation is shown in Figure 2. Respondents were generally engaged in the group deliberations; in the horizontal condition, 74-76% of group members participated more than once in each deliberation, and in the hierarchical condition, 76-82% of group members participated more than once in each deliberation, with leaders displaying 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 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.\(^6\)

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 verus all 100 servicepeople on board. Half of respondents in each experimental condition receive each treatment.

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 on-

---

\(^5\) Respondents in the group conditions deliberated using chat platform constructed in SMARTRIQS (Molnar, 2019).

\(^6\) 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.
going 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.

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 koreas 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 its an unfortunate accident. assuming nk did it will cause more panic.
unless another incident occur
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.
Advisor2: All service people survived. So is is unlikely that this was an attack
Leader: I agree advisor 2, we need more details to come to a informed decision
Advisor4: we can summer somewhat unlikely and just monitor any suspicious activities
Advisor1: search & 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 need more detailed investigation
Advisor3: Accusations could escalate quickly given their leader's nature.
*** Advisor1 has left the chat ***
Advisor3: I say it is unlikely.
*** Advisor4 has left the chat ***
*** Advisor3 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.

2.1 Hypotheses

Our core hypotheses 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.
To begin, we expect that, in individual decision-making contexts, we will replicate canonical cognitive biases that could lead decision-makers to support more “hawkish” solutions to political conflict:

H1. Hawkish Biases in Individual Decision Making:
- **H1a. Risk-Taking to Avoid Loss**: Individuals will be more risk-seeking in the domain of losses than the domain of gains.
- **H1b. Consequences & Assessments of Intentionality**: Individuals will be more likely to assess an incident as intentional when its costs are higher.
- **H1c. Devaluation of Proposals from Adversaries**: Individuals will evaluate a proposal from an adversary more negatively than the same proposal from their own side.

Given the competing findings from the extant literature on the efficacy of groups in reducing biases, however, the effect of groups could go in many directions – increasing, decreasing, or having no impact.

H2. The Relative Efficacy of Groups:
- **H2a. Benefit of Groups**: Group dialogues and debate may *reduce* the prevalence of biases, as compared to individual decision-making.
- **H2b. Risks of Groups**: Group pathologies such as the desire for conformity or group polarization may *exacerbate* biases, as compared to individual decision-making.
- **H2c. Negligible Impact of Groups**: The deeply ingrained, subconscious nature of most hawkish biases may cause groups to have *no* effect on the prevalence of these biases.

3 Analysis

To test these hypotheses, we first look within each group treatment (individual, horizontal, hierarchical) to examine the prevalence of each tested bias across the vignette treatments (gain/loss frame, 0/100 fatalities, US/China-authorship). We then compare these differences *across* groups to assess the extent to which these different decision-making structures affect susceptibility to each of the tested biases. Finally, we probe the
robustness of our findings, assessing the degree to which various types of leader characteristics or group diversity impacts susceptibility to biases and the ability to reach a decision in the first place.

3.1 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.

Figure 3: Prospect theory framing effects replicate in groups

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 decision rule. See Figure 4 for additional decision rule results.

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 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.\footnote{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 < 0.02, 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.}

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 \textit{unanimous} decision, we find similar results as in the hierarchical condition: the group setting \textit{increases} susceptibility to these framing effects (\( p < .004 \)). However, if we examine the full set of horizontal groups using a less stringent decision rule, such as a majority rule (\( p < .08 \)) 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.

### 3.2 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 (e.g. choose policy A or B), 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
Figure 4 displays the effect of the domain of losses on the probability of risky choice, within horizontal groups using different decision rules. The figure shows that the canonical prospect theory result replicates across all three types of horizontal decision rules (majority rule, median voter, unanimity rule), but is the largest in unanimous groups (significantly larger than in the individual condition: $p < 0.004$). Point estimates are cell means with 90% and 95% bootstrapped confidence intervals.
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).

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 decision rule. See Figure 6 for additional decision rule results.

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 more susceptible to framing effects, in this case groups perform similarly to individuals.8

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

8Comparing 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, $p < 0.84$ with controls at the leader-level, $p < 0.85$ with controls at the group-level. See Appendix §2.1.
similar extent as individuals do.

Figure 6: Intentionality bias effects by horizontal decision rule

Figure 6 displays the degree of perceived intentionality, given fatalities, within horizontal groups using different decision rules. The figure shows that the canonical intentionality bias result replicates across all three types of horizontal decision rules (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.

3.3 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 not replicate the standard reactive devaluation result across two of the three decision-making units. 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 (Brehm and Brehm, 2013) and reliance on source credibility as a heuristic for value (Hovland and Weiss, 1951). 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 (Maoz et al., 2002). 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 simply because the rivalry was less clear-cut than the violent, intractable conflicts in which this bias has historically been studied. In others words, Israelis may be more suspicious and distrusting of Palestinians (Maoz et al., 2002) and Americans more distrusting of the Soviet Union or North Vietnamese during the Cold War (Ross and Ward, 1995; Ashmore et al., 1979) than Americans today are of China, with whom the United States has a less directly confrontational relationship.

However, even with the specificity of this proposal and ambiguity of the rivalry, we do see the phenomenon of reactive devaluation replicating in horizontal groups, particularly those groups that reached a unanimous decision (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.
Figure 7: Reactive devaluation experiment displays mixed results

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 decision rule. See Figure 8 for additional decision rule results.
Figure 8: Reactive devaluation effects by horizontal decision rule

Figure 8 displays the effect of Chinese (vs American) authorship on support for the policy proposal within horizontal groups using different decision rules. The figure shows the canonical reactive devaluation result replicates in two of the three decision rules (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.
3.4 Extensions and limitations

Thus far, our results suggest that a trio of canonical biases in the judgment and decision-making literature – sensitivity to framing effects in prospect theory, intentionality bias, and reactive devaluation of adversary proposals — all persist or become even more pronounced in group settings. However, there are a number of important limitations and caveats worth discussing.

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 (Allison, 1971). 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.9 On the one hand, 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 (Janis, 1972). In that sense, the fact that we replicate these three biases 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? To test this question, we take advantage of the lengthy battery of individual differences administered

---

9Although 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.
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 (Ratkovic and Tingley, 2017) regressing our dependent variable on the treatment, a vector of 21
individual differences (foreign policy orientations, personality traits, demographic characteristics, histories
of leadership, and so on), and interactions between these leader-level traits and the treatment using data
from the hierarchical condition. 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.10

Third, 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 (Horowitz et al., 2019; Page, 2019). Diverse groups are thought to lead to more
extensive debate, increase exposure to other viewpoints, introduce differences in risk preferences, and
avoid group pathologies such as groupthink (Janis, 1972), where striving for uniformity may overwhelm
accuracy motives. Yet groups that are too diverse may move too far in the other direction, become so
divisive that they suffer from a polythink dynamic (Mintz and Wayne, 2016) in which they are unable to
reach consensus at all.

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-
sity in a multidimensional fashion, calculating the group-level variance of a given trait in each group, and
averaging across diversity scores for three types of traits, to produce measures of three different types of
diversity. We begin by considering groups whose members hold different ex ante attitudes about politics
– including right-wing authoritarianism, political ideology and social dominance orientation. Next, we

10It is of course possible that groups composed of actual elite decision-makers would display significantly different findings,
although a recent meta-analysis of paired experiments on elite and mass samples suggests some grounds for skepticism (Kertzer,
Forthcoming). Moreover, we urge caution to scholars seeking to field similar studies on elite samples: studies on real foreign policy
decision-makers invariably involve smaller sample sizes – effectively made smaller still once analyzed at the group-level – raising
concerns about statistical power.
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’ attitudes, demographic characteristics, or prior experiences). 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.
examine diversity from a demographic perspective, whereby more diverse groups are those whose members come from different socio-economic, racial, and religious backgrounds and include a mixture of men and women. Finally, we explore diversity of experience within groups, where different members of the group have varying levels of past experience in leadership (political or otherwise).

Figure 10: More diverse groups more likely to experience dissension

Regardless of how we operationalize diversity, however, we find no effects of diversity on susceptibility to any of the hawkish biases we examine. As Figure 9 shows, diverse groups are 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.\(^{11}\) It is not the case, however, that diversity displays no effects whatsoever: as Figure 10 shows, we find that more diverse groups – particularly those with more diverse 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

\(^{11}\)As 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 doesn’t significantly interact with the treatment regardless of the functional form used, while the LASSOplus results above suggest that leader gender does not significantly affect group decisions in the hierarchical condition either.
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.\textsuperscript{12} Nonetheless, it does not appear to be the case that more diverse groups are less likely to display these three tendencies.

Finally, there are two other alternative caveats of the results worth noting, which also serve as alternative interpretations for our results. One is that for the ameliorative effects of aggregation to take place, group members need to interact face-to-face (Holmes, 2018) rather than deliberate at a distance. 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" (e.g. Surowiecki 2005) might suggest greater rationality as groups grow in size. 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 (for a review, see Jordan et al. 2009). John F. Kennedy, for example, was disappointed by the results of a large number of advisors as it related to the failed Bay of Pigs invasion. “The advice of every member of the Executive Branch brought in to advise was unanimous - and the advice was wrong.” In response 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

\textsuperscript{12}The 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.
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 (Holmes, 2018), we can test this informational mechanism directly by testing whether groups where respondents exchanged more information with one another as part of their deliberations displayed weaker biases than groups where respondents communicated less. 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{13}

One reason 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 (Walther, 1992). 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 (Antheunis, Valkenburg and Peter, 2007; Tidwell and Walther, 2006). For example, Wheeler and Holmes (2021) 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. 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.

### 4 Conclusion

In a recent review of the problem of aggregation, Gildea (Forthcoming, 1-2) 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." To date such

\textsuperscript{13}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.
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 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 need not 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" Johnson (2015, 760), 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 (1998) applied the bias to a number of cases, focusing "on a unitary actor embodied by the president" (p. 187), 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 (1990, 58) 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 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, 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 (Brutger and Kertzer, 2018), 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.

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. 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 (Kertzer, 2016; Stein, 2017; Mintz, Valentino and Wayne, 2022). 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 biases, they introduce new dynamics that may exacerbate deviations from expected utility models. Early psychological research on group conformity (Sherif, 1935; Asch and Guetzkow, 1951; Janis, 1972) 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 (e.g. Badie 2010). It may be, however, that groupthink is receiving unfair blame. As Whyte (1989, 40) has argued, "history and the daily newspaper provide examples of policy decisions made by groups that re-
sulted 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. 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 (McDermott, 2004; Sheffer et al., 2018), intentionality bias is central to questions of blame attribution in politics more generally (McGraw, 1991; Malhotra and Kuo, 2008), and reactive devaluation is tightly linked to our theories of negative partisanship (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 micro dyadic interactions between individual diplomats, as well as collective decision-making in which diplomats conform with logics of practice or habit (McCourt, 2016). During NATO decision-making sessions regarding the proposed use of force in Libya in 2011, for example, Adler-Nissen and Pouliot (2014) 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 (Ringmar, 2014, 6), 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.
References


36


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


Hawkish Biases & Group Decision-Making:
Supplementary Appendix

Contents

1 Experimental Instrumentation ............................................. A1
  1.1 Sensitivity to Gain/Loss Framing .................................. A1
  1.2 Intentionality Bias ..................................................... A1
  1.3 Reactive Devaluation ................................................... A1

2 Survey Flow ................................................................. A3
  2.1 Survey Flow and Attrition ............................................. A3
     Table 2.2: Prospect theory experiment .......................... A5
     Table 2.3: Intentionality bias experiment ........................ A6
     Table 2.4: Reactive devaluation experiment ..................... A7
  2.2 Sensitivity analyses .................................................... A8
     Table 2.5: Prospect theory sensitivity analysis ................. A10

3 Group size analysis and simulations ...................................... A10
  Table 2.6: Intentionality bias sensitivity analysis ............... A11
  Table 2.7: Reactive devaluation sensitivity analysis .......... A12
  Figure 3.2: Hawkish biases do not aggregate in larger groups A14

4 Group participation analysis .............................................. A15
  Table 4.8: Little evidence of heterogeneous effects by participation A17
1 Experimental Instrumentation

1.1 Sensitivity to Gain/Loss Framing

LOSS FRAME: 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:

- Policy A: 400 people will die
- Policy B: There is a $\frac{1}{3}$ probability that nobody will die, and a $\frac{2}{3}$ probability that 600 people will die.

GAIN FRAME: 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:

- Policy A: 200 people will be saved
- Policy B: There is a $\frac{1}{3}$ probability that 600 people will be saved, and a $\frac{2}{3}$ probability that no people will be saved

1.2 Intentionality Bias

NO FATALITIES:
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. Fortunately, there were no fatalities as all servicepeople on the boat were rescued.

ALL FATALITIES:
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. Unfortunately, there were 100 fatalities as none of the servicepeople on the boat could be rescued.

1.3 Reactive Devaluation

CHINA AUTHORSHIP:
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 government submitted a brief proposal to the United States government containing their conditions for continuing higher-level talks on the core
issues (including tariffs, currency manipulation, and intellectual property). The main components of this proposal are listed below:

1. China will remove up to 50% of the new tariffs introduced since January 2018, in exchange for parallel removal of tariffs imposed by the US government.

2. The US Department of Treasury will remove their recent designation of China as a currency manipulator in exchange for China increasing the value of the Chinese Yuan back to 2018 levels.

3. The mutual opening of new areas of domestic commerce to foreign direct investment, including reducing regulations that currently mandate foreign companies to transfer technology as a condition for securing investment approvals.

4. The establishment of a new UN watchdog agency specifically responsible for ensuring the protection of US and Chinese intellectual property and patent rights.

Collectively, China hopes that these measures will reduce current tensions and ensure a productive, mutually beneficial trade relationship moving forward.

US AUTHORSHIP:

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 United States government submitted a brief proposal to the Chinese government containing their conditions for continuing higher-level talks on the core issues (including tariffs, currency manipulation, and intellectual property). The main components of this proposal are listed below:

1. The United States will remove up to 50% of the new tariffs introduced since January 2018, in exchange for parallel removal of tariffs imposed by the Chinese government.

2. The Central Bank of China will increase the value of the Chinese Yuan back to 2018 levels in exchange for the US Department of Treasury’s removing their recent designation of China as a currency manipulator.

3. The mutual opening of new areas of domestic commerce to foreign direct investment, including reducing regulations that currently mandate foreign companies to transfer technology as a condition for securing investment approvals.

4. The establishment of a new UN watchdog agency specifically responsible for ensuring the protection of US and Chinese intellectual property and patent rights.

Collectively, the United States hopes that these measures will reduce current tensions and ensure a pro-
ductive, mutually beneficial trade relationship moving forward.

2 Survey Flow

2.1 Survey Flow and Attrition

In this paper we test a series of hypotheses utilizing experiments (prospect theory, intentionality bias, and reactive devaluation) in three conditions: an individual condition, horizontal group, and hierarchical group. The latter two conditions utilized an online software package, SMARTRIQS, to create real-time respondent interaction environments in Qualtrics (Molnar, 2019). After completing an individual difference and demographic battery and passing a set of attention checks, respondents were randomly assigned to one of these conditions. In the individual condition respondents were presented with the three experiments and dependent variables of interest measured. In the two group conditions, respondents were paired with other respondents after completing the demographics but before being presented with the experimental vignettes.

There are several points worth mentioning with respect to the design of the interactive environment created. First, respondents were only able to move on to the experimental modules once there were five respondents successfully paired, forming a full group. To accomplish this, we built a waiting room in SMARTRIQS, where respondents are held until five are present. Once a full group formed, they would move on together to the first experiment.

After each module, respondents were placed in a similar waiting room while the respondents in the group finished completing their responses. Unlike the first waiting room. These waiting rooms were not used to pair new participants. Instead, these subsequent waiting rooms were used to keep members of the same group in sync throughout the study.

The initial matching wait time (for pairing participants initially) lasted a maximum of five minutes. If subjects were not able to be matched within that five minutes the survey was terminated, and subjects did not complete any of the modules.

Subsequent, post-matching waiting rooms (for keeping group members in sync) were limited to a four minute wait time. If all members of a group did not arrive at a wait room before time was up (perhaps because a group member had dropped out due to a faulty connection) the group was allowed to continue on to the next module without the missing group members.

\footnote{On the virtue of experimental methods in political science more generally, see McDermott (2002).}
As we discuss below, if groups complete a module with fewer than three members (or in the hierarchical condition, if a leader was not present), we did not include that group in our main analysis, although we employ a variety of robustness checks below to show that our results are not sensitive to these inclusion criteria.

Prior to each group decision group members deliberated in an online chatroom. In the chatroom each group member had a unique identifier. In the Hierarchical condition, the identifiers were: Leader, Advisor-1, Advisor-2, Advisor-3, Advisor-4. In the Horizontal condition, the identifiers were: Group-member-1, Group-member-2, Group-member-3, Group-member-4, Group-member-5. Participants were notified of their identifier prior to each chat, and there was a reminder below the chat box (so that participants would not forget who they were in the chat). The group chatroom included a timer that identified how much time remained. Once the allotted time was up, respondents would all move to the next screen, typically one where dependent variables of interest were measured.

Because subjects in the group conditions had to interact in real time, there were several points in the survey where subjects could drop out. There are three types of attrition or missingness one might be concerned about, each of which we investigate in detail below.

First, if five respondents were not present at initial matching then the experimental section did not begin and therefore the survey automatically ended for these respondents having only completed the demographics section. Attrition prior to or during initial matching is the least problematic type of attrition for our purposes, as this type of attrition happens before any of the treatments are administered. Therefore, attrition at this stage cannot introduce post-treatment bias, in that exposure to the submodule-level treatment itself could not have caused subjects to drop from our study. However attrition at this stage could still lead to compositional differences between participants in the group conditions (who had to be matched) vs participants in the individual condition (who did not have to be matched with other participants). We guard against this threat by statistically controlling for a wide array of pre-treatment variables, including demographic characteristics (age, gender, education, income, religion, race), political attitudes (partisanship, interest in politics, and foreign policy orientations like hawkishness and isolationism (Kertzer et al., 2014)), and a plethora of individual differences from political psychology (including need for cognition (Cacioppo and Petty, 1982), social dominance orientation (Pratto et al., 1994), right-wing authoritarianism (Altemeyer, 1981), aggression (Buss and Perry, 1992), risk attitudes (Kertzer, 2017), and the “big 5” personality traits (McCrae and Costa, 1987)). Crucially, as the results in Tables 2.2-2.4 show, our results hold both with and without including these controls.
## Table 2.2: Prospect theory experiment

<table>
<thead>
<tr>
<th></th>
<th>Individual vs Horizontal</th>
<th>Individual vs Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) (2) (3) (4) (5) (6) (7)</td>
<td></td>
</tr>
<tr>
<td>Loss frame</td>
<td>0.329*** 0.323*** 0.324*** 0.324*** 0.329*** 0.322*** 0.324***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.033) (0.033) (0.033) (0.033) (0.033) (0.033) (0.033)</td>
<td></td>
</tr>
<tr>
<td>Horizontal condition</td>
<td>−0.018 −0.017 −0.058 −0.020</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.040) (0.041) (0.053) (0.041)</td>
<td></td>
</tr>
<tr>
<td>Loss x Horizontal</td>
<td>0.068 0.079 0.198*** 0.098*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.056) (0.056) (0.070) (0.057)</td>
<td></td>
</tr>
<tr>
<td>Hierarchical condition</td>
<td>−0.117*** −0.120*** −0.120***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.042) (0.043) (0.043)</td>
<td></td>
</tr>
<tr>
<td>Loss x Hierarchical</td>
<td>0.438*** 0.483*** 0.483*** 0.487***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.023) (0.159) (0.162) (0.160)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.438*** 0.483*** 0.483*** 0.487***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.023) (0.140) (0.158) (0.158)</td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>No Yes Yes No</td>
<td></td>
</tr>
<tr>
<td>Horizontal decision rule</td>
<td>Median voter Median voter Unanimity rule Majority rule</td>
<td></td>
</tr>
<tr>
<td>Hierarchical controls</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1,164 1,164 979 1,146 1,127 1,127 1,127</td>
<td></td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.133 0.136 0.155 0.140 0.160 0.161 0.162</td>
<td></td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01. Models 2-4 and 6-7 also control for age, gender, education, income, religion, race, party ID, political interest, need for cognition, SDO, RWA, aggression, risk attitudes, militant internationalism, isolationism, extraversion, agreeableness, conscientiousness, neuroticism, openness.
<table>
<thead>
<tr>
<th></th>
<th>Individual vs Horizontal</th>
<th>Individual vs Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Fatalities</td>
<td>0.097***</td>
<td>0.103***</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Horizontal condition</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Fatalities x Horizontal</td>
<td>0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>0.012</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Fatalities x Hierarchical</td>
<td>-0.005</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.592***</td>
<td>0.547***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.079)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Horizontal decision rule</td>
<td>Median voter</td>
<td>Median voter</td>
</tr>
<tr>
<td></td>
<td>1,044</td>
<td>1,044</td>
</tr>
<tr>
<td>Hierarchical controls</td>
<td>Leader</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>0.042</td>
<td>0.110</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.117</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>0.103</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>0.095</td>
<td></td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01. Models 2-4 and 6-7 also control for age, gender, education, income, religion, race, party ID, political interest, need for cognition, SDO, RWA, aggression, risk attitudes, militant internationalism, isolationism, extraversion, agreeableness, conscientiousness, neuroticism, openness.
Table 2.4: Reactive devaluation experiment

<table>
<thead>
<tr>
<th></th>
<th>Individual vs Horizontal</th>
<th>Individual vs Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>China authored</td>
<td>-0.019</td>
<td>-0.019</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Horizontal condition</td>
<td>0.064***</td>
<td>0.059**</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>China x Horizontal</td>
<td>-0.045</td>
<td>-0.049</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>-0.008</td>
<td>-0.019</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>China x Hierarchical</td>
<td>0.032</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.625***</td>
<td>0.577***</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Horizontal decision</td>
<td>Median</td>
<td>Median</td>
</tr>
<tr>
<td>rule</td>
<td>voter</td>
<td>voter</td>
</tr>
<tr>
<td>N</td>
<td>988</td>
<td>988</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.010</td>
<td>0.057</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01. Models 2-4 and 6-7 also control for age, gender, education, income, religion, race, party ID, political interest, need for cognition, SDO, RWA, aggression, risk attitudes, militant internationalism, isolationism, extraversion, agreeableness, conscientiousness, neuroticism, openness.
Second, participants might drop out during or between each of the three modules. Attrition at these stages occurred after subjects had been treated with one or more of our experimental conditions. This raises the possibility of post-treatment bias. That is, one of the between-group treatments could have caused differential attrition, which might then cause differences in the distribution of potential outcomes across subsequent treatments. While we cannot exclude this possibility entirely, we can run a variety of sensitivity analyses to ascertain the robustness of our findings. This is also particularly important given the possibility that differential attrition occurred along an unmeasured dimension that cannot be controlled for in the regression analyses in Tables 2.2-2.4.

Finally, there is a third and related type of attrition that can occur, which we also incorporate into the sensitivity analyses below, which refers not to respondents who drop out from the study, but respondents who are removed from the sample by the experimenters. Our analyses in the main text exclude respondents who complete a given experiment in a group below a certain size (in the horizontal condition, groups must have at least three members, and in the hierarchical condition, groups must have at least three members, one of whom is also the leader of the group). These criteria are included both for reasons of construct validity (a hierarchical group is no longer hierarchical if it doesn’t have a leader), and because it makes our results in the main text a more conservative test (if we’re interested in investigating whether group interaction causes hawkish biases to dissipate, we want the groups to be larger than just two members). However, these criteria might also raise concerns if the types of groups that meet these criteria are filled with systematically different types of people than in the individual condition. Therefore, we also relax these criteria as part of our sensitivity analyses below.²

2.2 Sensitivity analyses

To assess the impact of these different types of attrition on our results, we therefore adopt a three-step sensitivity analysis.

- First, we re-estimate our analysis from the main text, but dropping the group size exclusion criteria (i.e. including horizontal and hierarchical groups that have fewer than three members; hierarchical groups are still only included in this analysis if one of the members is the group leader). These complete observation results, shown in models 5 and 11 of Tables 2.5-2.7, find strikingly similar patterns as the results depicted in the main text (reproduced here in models 4 and 10).

²There is one additional exclusion criterion in the studies, which we preserve in the analyses below: we remove any group from the analysis where one respondent was flagged as a “bot”; since bots produce random responses, inclusion of bot responses in our analysis adds noise but does not substantively change our results.
• Second, we then conduct an “extreme bounds” analysis, similar in spirit to Manski bounds (Manski, 1990). This analysis, shown in models 2-3, 6-7 and 12-13 of Tables 2.5-2.7, replaces missing values with maximum or minimum possible values on the dependent variable to get bounds on how biased our data could possibly be due to non-random missingness. As its name suggests, extreme bounds analysis is necessarily conservative, since replacing missing values with maximum or minimum possible values creates relatively large confidence intervals around the true effect. Nonetheless, we obtain strikingly similar results as in our base models, consistently replicating the prospect theory and intentionality bias results in all three group conditions, and the reactive devaluation result in just the horizontal condition, as was the case in the main text.

• Third, to conduct a more theoretically motivated sensitivity test, we use a machine learning approach. We fit a neural net on the individual data as the training set, using our treatment condition and extensive pre-treatment demographic battery as predictors, tuning the model using a cross-validation approach (Kuhn and Johnson, 2013). We then use that model to create predicted values for the missing responses in the horizontal and hierarchical data using the treatment conditions and demographic covariates observed for those respondents. We then consider two test cases.

  – First, what would happen if our missing respondents were “antisocial”, behaving in the group condition exactly as the neural net suggests they would have in the individual condition, rather than being influenced by other members of the group? To assess the stability of our findings in this case, in the horizontal condition we calculate the new median vote for each group, taking the input of the predicted votes from the neural net into account (model 8 in Tables 2.5-2.7). For hierarchical groups with missing leaders, we similarly replace the missing leader votes with these predicted votes (model 14 in Tables 2.5-2.7). In both cases, we obtain the same findings as in the main text.

  – Second, what would happen if our missing respondents were “influencers”, the most persuasive members of their groups, convincing others to come on board? To estimate the results for this case, in the horizontal condition, we treat the missing group member as the pivotal vote (model

---

3For a similar extreme bound application in political science, see Margalit (2021).

4Our approach differs somewhat from traditional Manski bounds because of the group-level structure of the data in the group conditions; we thus adopt a two-stage strategy. For horizontal groups we first replace missing values for each respondent (setting them to either the maximum or the minimum value), and then use the same median-voter rule from the main text to calculate the decision. For hierarchical groups, the approach is simpler, since the hierarchical group decision is equivalent to the imputed leader choice by design.
As before, we replicate the results from the main text.

Altogether, this analysis suggests that even under relatively extreme assumptions, we still find evidence in support of our central finding in the manuscript, which is that hawkish biases do not aggregate away in group contexts.

Table 2.5: Prospect theory sensitivity analysis

<table>
<thead>
<tr>
<th></th>
<th>Individual</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>Loss Trt</td>
<td>0.329**</td>
<td>0.329**</td>
<td>0.329**</td>
<td>0.398***</td>
<td>0.350***</td>
<td>0.348***</td>
<td>0.347***</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.033)</td>
<td>(0.033)</td>
<td>(0.043)</td>
<td>(0.032)</td>
<td>(0.032)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.438***</td>
<td>0.438***</td>
<td>0.438***</td>
<td>0.421***</td>
<td>0.435***</td>
<td>0.431***</td>
<td>0.441***</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.024)</td>
<td>(0.024)</td>
<td>(0.031)</td>
<td>(0.023)</td>
<td>(0.023)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Model</td>
<td>Base</td>
<td>Extreme</td>
<td>Extreme</td>
<td>Base</td>
<td>Complete</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td>N</td>
<td>760</td>
<td>760</td>
<td>760</td>
<td>404</td>
<td>735</td>
<td>736</td>
<td>736</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.112</td>
<td>0.112</td>
<td>0.112</td>
<td>0.175</td>
<td>0.138</td>
<td>0.136</td>
<td>0.136</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Horizontal</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(8)</td>
<td>(9)</td>
<td>(10)</td>
<td>(11)</td>
<td>(12)</td>
<td>(13)</td>
<td>(14)</td>
</tr>
<tr>
<td>Loss Trt</td>
<td>0.354***</td>
<td>0.354***</td>
<td>0.506***</td>
<td>0.439***</td>
<td>0.435***</td>
<td>0.432***</td>
<td>0.440***</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.032)</td>
<td>(0.044)</td>
<td>(0.036)</td>
<td>(0.036)</td>
<td>(0.036)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.434***</td>
<td>0.434***</td>
<td>0.321***</td>
<td>0.366***</td>
<td>0.361***</td>
<td>0.376***</td>
<td>0.368***</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.023)</td>
<td>(0.033)</td>
<td>(0.027)</td>
<td>(0.027)</td>
<td>(0.027)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>Model</td>
<td>Neural Net</td>
<td>Neural Net</td>
<td>Base</td>
<td>Complete</td>
<td>Extreme</td>
<td>Extreme</td>
<td>Neural net</td>
</tr>
<tr>
<td>N</td>
<td>736</td>
<td>736</td>
<td>367</td>
<td>581</td>
<td>589</td>
<td>589</td>
<td>589</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.141</td>
<td>0.141</td>
<td>0.261</td>
<td>0.199</td>
<td>0.194</td>
<td>0.194</td>
<td>0.200</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01

3 Group size analysis and simulations

One potential interpretation of our findings is that we replicate hawkish biases in group decision-making contexts because our groups are too small for the “miracle of aggregation” to kick in. There are perhaps two potential implications of this argument, each of which we test here. First, if hawkish biases decrease with group size, we should expect that group size display a negative interaction effect with the
### Table 2.6: Intentionality bias sensitivity analysis

<table>
<thead>
<tr>
<th></th>
<th>Individual</th>
<th>Horizontally</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>Fatal Trt</td>
<td>0.097***</td>
<td>0.088***</td>
<td>0.099***</td>
<td>0.099***</td>
<td>0.075***</td>
<td>0.044*</td>
<td>0.071***</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.018)</td>
<td>(0.017)</td>
<td>(0.026)</td>
<td>(0.019)</td>
<td>(0.025)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.592***</td>
<td>0.586***</td>
<td>0.597***</td>
<td>0.609***</td>
<td>0.620***</td>
<td>0.467***</td>
<td>0.714***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.018)</td>
<td>(0.014)</td>
<td>(0.018)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Model</td>
<td>Base</td>
<td>Extreme</td>
<td>Extreme</td>
<td>Base</td>
<td>Complete</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td></td>
<td>Obs</td>
<td>Min</td>
<td>Max</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>748</td>
<td>760</td>
<td>760</td>
<td>296</td>
<td>600</td>
<td>722</td>
<td>722</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.041</td>
<td>0.031</td>
<td>0.042</td>
<td>0.045</td>
<td>0.022</td>
<td>0.003</td>
<td>0.020</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Horizontal</th>
<th>Hierarchical</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(8)</td>
<td>(9)</td>
<td>(10)</td>
<td>(11)</td>
<td>(12)</td>
<td>(13)</td>
<td>(14)</td>
</tr>
<tr>
<td>Fatal Trt</td>
<td>0.083***</td>
<td>0.077***</td>
<td>0.092***</td>
<td>0.080***</td>
<td>0.050*</td>
<td>0.072***</td>
<td>0.085***</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.014)</td>
<td>(0.025)</td>
<td>(0.023)</td>
<td>(0.028)</td>
<td>(0.022)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.614***</td>
<td>0.618***</td>
<td>0.604***</td>
<td>0.610***</td>
<td>0.498***</td>
<td>0.682***</td>
<td>0.607***</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.010)</td>
<td>(0.018)</td>
<td>(0.016)</td>
<td>(0.020)</td>
<td>(0.015)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Model</td>
<td>Neural Net</td>
<td>Neural Net</td>
<td>Base</td>
<td>Complete</td>
<td>Extreme</td>
<td>Extreme</td>
<td>Neural Net</td>
</tr>
<tr>
<td></td>
<td>Antisocial</td>
<td>Influencer</td>
<td>Obs</td>
<td>Min</td>
<td>Max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>722</td>
<td>722</td>
<td>365</td>
<td>474</td>
<td>589</td>
<td>589</td>
<td>589</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.041</td>
<td>0.042</td>
<td>0.032</td>
<td>0.023</td>
<td>0.004</td>
<td>0.017</td>
<td>0.033</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01
Table 2.7: Reactive devaluation sensitivity analysis

<table>
<thead>
<tr>
<th></th>
<th>Individual</th>
<th></th>
<th>Horizontal</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>China Trt</td>
<td>−0.019</td>
<td>−0.021</td>
<td>−0.016</td>
<td>−0.063**</td>
<td>−0.089***</td>
<td>−0.087***</td>
<td>−0.045**</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.018)</td>
<td>(0.017)</td>
<td>(0.020)</td>
<td>(0.022)</td>
<td>(0.027)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.625***</td>
<td>0.603***</td>
<td>0.637***</td>
<td>0.687***</td>
<td>0.700***</td>
<td>0.483***</td>
<td>0.797***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.014)</td>
<td>(0.013)</td>
<td>(0.015)</td>
<td>(0.019)</td>
<td>(0.019)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Model</td>
<td>Base</td>
<td>Extreme</td>
<td>Extreme</td>
<td>Base</td>
<td>Complete</td>
<td>Extreme</td>
<td>Extreme</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Min</td>
<td>Max</td>
<td>Min</td>
<td>Max</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>7732</td>
<td>790</td>
<td>790</td>
<td>258</td>
<td>540</td>
<td>706</td>
<td>706</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.0003</td>
<td>0.0005</td>
<td>−0.0002</td>
<td>0.013</td>
<td>0.028</td>
<td>0.013</td>
<td>0.007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Horizontal</th>
<th></th>
<th>Hierarchical</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(8)</td>
<td>(9)</td>
<td>(10)</td>
<td>(11)</td>
<td>(12)</td>
<td>(13)</td>
<td>(14)</td>
</tr>
<tr>
<td>China Trt</td>
<td>−0.065***</td>
<td>−0.043***</td>
<td>0.013</td>
<td>−0.004</td>
<td>0.005</td>
<td>−0.008</td>
<td>−0.004</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.013)</td>
<td>(0.028)</td>
<td>(0.024)</td>
<td>(0.029)</td>
<td>(0.023)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.678***</td>
<td>0.662***</td>
<td>0.617***</td>
<td>0.620***</td>
<td>0.452***</td>
<td>0.723***</td>
<td>0.620***</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.009)</td>
<td>(0.019)</td>
<td>(0.017)</td>
<td>(0.020)</td>
<td>(0.015)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Model</td>
<td>Neural Net</td>
<td>Neural Net</td>
<td>Base</td>
<td>Complete</td>
<td>Extreme</td>
<td>Extreme</td>
<td>Neural Net</td>
</tr>
<tr>
<td></td>
<td>Antisocial</td>
<td>Influencer</td>
<td>Obs</td>
<td>Min</td>
<td>Max</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>N</td>
<td>706</td>
<td>706</td>
<td>343</td>
<td>433</td>
<td>589</td>
<td>589</td>
<td>589</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.024</td>
<td>0.013</td>
<td>−0.002</td>
<td>−0.002</td>
<td>−0.002</td>
<td>−0.002</td>
<td>−0.002</td>
</tr>
</tbody>
</table>

*p < .1; **p < .05; ***p < .01
experimental treatments. We therefore estimate a series of regression models interacting each study-level treatment with group size, for both horizontal groups, and hierarchical groups. Crucially, in no case are the interaction effects statistically significant.

Second, it is possible that even our groups of five are too small. If this were the case it would be an important limitation on the applicability of the miracle of aggregation for foreign policy decision-making, since the “inner circle” in many foreign policy decision-making groups can often be rather small (Jost, 2021). Nonetheless, we investigate this claim using a simulation-based approach building on LeVeck and Narang (2017).

- Randomly assign treatment conditions to 1000 groups (e.g. \texttt{rbinom(1000,1,0.5)})
- Populate each group with \(N\) respondents, sampling with replacement from the observations in the individual condition (but only sampling from the same experiment-level treatment condition within each group).
- Use the median voter rule to determine each group decision.
- Use the data from these 1000 groups to calculate the ATE.
- Repeat the above, resampling \(B = 1500\) times.
- Repeat the above for all group sizes, varying \(N\) from 1 to 20.

The results of the simulations, plotted in Figure 3.2, fail to support the claim that hawkish biases aggregate away in larger groups. For the intentionality bias and reactive devaluation experiments, the effect magnitude stays fairly constant as we increase the size of our simulated groups (although the reactive devaluation effect becomes statistically significant once we increase our group size). For the prospect theory experiment we find a sharp \textit{increase} in the magnitude of the effect as we increase our simulated group size, consistent with other research on group decision-making that has found that social influence and voting over binary alternatives amplifies the majority opinion regardless of its accuracy (Becker, Guilbeault and Smith, 2018).
Figure 3.2: Hawkish biases do not aggregate away in larger groups

Each panel displays the results of the simulations from each experimental module, where we resample data from respondents in the individual condition to create artificial groups of sizes ranging from 1 – 20, imputing the group decision using the median voter rule. These simulations suggest that hawkish biases do not aggregate away in larger groups: if anything, the treatment effect in the prospect theory experiment increases logarithmically with simulated group size.
4 Group participation analysis

As noted in the main text, we find evidence that a trio of biases from the judgment and decision-making literature that scholars have deemed relevant for foreign policy — risk-taking in the domain of losses, intentionality bias, and reactive devaluation — replicate in group decision-making contexts, and are sometimes even larger in group contexts than individual ones.

One potential question that these findings provoke is whether our results are simply due to respondents in the group condition are not taking the study seriously. We have a number of reasons to believe this is not the case. First, as noted in Appendix §2.1, respondents in all conditions had to pass a serious of attention checks in order to make it to the experimental modules, such that the least attentive respondents would have been excluded prior to randomization. Second, as noted in the main paper, an analysis of the deliberation transcripts (an example of which is shown in Figure 2 in the main text) shows respondents were relatively engaged in the group deliberations: in the horizontal condition, 74% of group members participate more than once in the deliberation in the prospect theory experiment, 75% do so in the intentionality bias experiment, and 74% in the reactive devaluation experiment. In the hierarchical condition, 86% of group leaders and 80% of advisers participate more than once in the deliberation in the prospect theory experiment, 82% of leaders and 76% of advisers do so in the intentionality bias experiment, and 77% of leaders and 75% of advisers do so in the reactive devaluation experiment.

Perhaps most importantly, we can exploit variation in participant engagement in our group conditions, to test empirically whether groups that featured greater rates of participation display weaker treatment effects than groups that featured lower rates of participation. One concern might be that absolute levels of group participation are confounded with group size, since groups with more members will also feature more participation; while we know from the analyses in Appendix §3 that our biases do not significantly vary with group size, we nonetheless sidestep this concern by calculating per capita participation rates per group. We then interact this group participation rate with the study-level treatments, and present the results in Table 4.8, which also conducts an analogous test in the individual condition by seeing whether individuals who wrote more in their deliberations display significantly larger treatment effects.

As Table 4.8 shows, we find no evidence that either horizontal or hierarchical groups that featured higher per-capita participation in the deliberation sessions displayed significantly smaller treatment effects than groups that featured lower per-capita participation. In the individual condition, we find that individuals who made more extensive justifications in the intentionality bias condition tended to be less
Table 4.8: Little evidence of heterogeneous effects by participation

<table>
<thead>
<tr>
<th>Prospect theory</th>
<th>Intentionality bias</th>
<th>Reactive devaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>0.420***</td>
<td>0.644***</td>
</tr>
<tr>
<td></td>
<td>(0.095)</td>
<td>(0.113)</td>
</tr>
<tr>
<td>Group participation</td>
<td>-0.008</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Trt x Group participation</td>
<td>-0.007</td>
<td>-0.043</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>Individual participation</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Trt x Indiv. participation</td>
<td>-0.0004</td>
<td>0.002**</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.444***</td>
<td>0.287***</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.083)</td>
</tr>
<tr>
<td>N</td>
<td>404</td>
<td>367</td>
</tr>
</tbody>
</table>

Adjusted $R^2$: 0.173, 0.262, 0.110

*p < .1; **p < .05; ***p < .01
responsive to the the fatalities treatment. Altogether, then, these results suggest little evidence that variation in participant engagement in the group condition is associated with different group decisions.
References


