Hawkish Biases & Group Decision-Making:
Supplementary Appendix

Contents

1 Study 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 1/3 probability that nobody will die, and a 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 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved

Outcome measures

INDIVIDUAL CONDITION:

What did you decide was the best choice?

- Policy A
- Policy B

HORIZONTAL CONDITION:

Did the group come to a unanimous decision?

- Yes
- No

[If yes]

What did the group decide was the best choice?

- Policy A
- Policy B

[If no]

What did you personally decide was the best choice?

- Policy A

- Policy B

HIERARCHICAL CONDITION:

[If leader]

As the group leader, what did you decide was the best choice?

- Policy A
- Policy B

[If advisor]

Though you are not the leader of the group, we are still interested in what you personally thought was the best choice.

- Policy A
- Policy B

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.

Outcome measures

INDIVIDUAL CONDITION:

How likely did you think it was that the vessel was *intentionally* sunk?

1. Extremely likely
2. Likely
3. Somewhat likely
4. Neither likely nor unlikely
5. Somewhat unlikely
6. Unlikely
7. Extremely unlikely

HORIZONTAL CONDITION:

Did your group come to a unanimous decision?

- Yes
- No

[If yes]

How likely did the group think it was that the vessel was *intentionally* sunk?

1. Extremely likely
2. Likely
3. Somewhat likely
4. Neither likely nor unlikely

5. Somewhat unlikely
6. Unlikely
7. Extremely unlikely

[If no]

How likely did you personally think it was that the vessel was *intentionally* sunk?

1. Extremely likely
2. Likely
3. Somewhat likely
4. Neither likely nor unlikely
5. Somewhat unlikely
6. Unlikely
7. Extremely unlikely

HIERARCHICAL CONDITION:

[If leader]

How likely did you think it was that the vessel was *intentionally* sunk?

1. Extremely likely
2. Likely
3. Somewhat likely
4. Neither likely nor unlikely
5. Somewhat unlikely
6. Unlikely
7. Extremely unlikely

[If advisor]

Even though you are not the leader, we are still interested in your views. How likely did you think it was that the vessel was *intentionally* sunk?

1. Extremely likely
2. Likely
3. Somewhat likely
4. Neither likely nor unlikely
5. Somewhat unlikely
6. Unlikely

7. Extremely unlikely

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 productive, mutually beneficial trade relationship moving forward.

Outcome measures

INDIVIDUAL CONDITION:

How much do you support the proposal, on a scale from 1 to 7?

1. Completely oppose
2.
3.
4.
5.
6.
7. Completely support

HORIZONTAL CONDITION:

Did your group come to a unanimous decision?

- Yes
- No

[If yes]

How much did the group support the proposal, on a scale from 1 to 7?

1. Completely oppose
2.
3.
4.
5.
6.
7. Completely support

[If no]

How much did you personally support the proposal, on a scale from 1 to 7?

1. Completely oppose
2.
3.
4.
5.
6.
7. Completely support

HIERARCHICAL CONDITION:

[If leader]

As the leader of your group, how much do you support the proposal, on a scale from 1 to 7?

1. Completely oppose
2.
3.
4.
5.
6.
7. Completely support

[If advisor]

While you aren't the leader, we are still interested in what you would have decided. How much do you support the proposal, on a scale from 1 to 7?

1. Completely oppose
2.
3.
4.
5.
6.
7. Completely support

1.4 Measures of Group Diversity

Diversity in Political Attitudes

To explore the impact of attitudinal diversity in the decision-making unit, we constructed a measure that accounts for variance in the group on various traits of interest, including the following traits:[1]

1. Political ideology (7-point Likert scale from extremely liberal to extremely conservative)
2. Military assertiveness (Herrmann, Tetlock and Visser, 1999; Kertzer et al., 2014)
    - The best way to ensure world peace is through American military strength
    - The use of military force only makes problems worse (R)
    - Rather than simply reacting to our enemies, it's better for us to strike first
3. Isolationism (Kertzer et al., 2014)
    - Generally, the more influence America has on other nations, the better off they are (R)
    - The U.S. needs to play an active role in solving conflicts around the world (R)
    - The U.S. government should just try to take care of the well-being of Americans and not get involved with other nations
4. Social dominance orientation (Ho et al., 2015)
    - An ideal society requires some groups to be on top and others to be on the bottom
    - Some groups of people are simply inferior to other groups
    - No one group should dominate in society (R)
    - Groups at the bottom are just as deserving as groups at the top (R)
    - We should do what we can to equalize conditions for different groups (R)
    - We should work to give all groups an equal chance to succeed (R)
5. Right wing authoritarianism (Feldman and Stenner, 1997)
    - Which one do you think is more important for a child to have: independence or respect for elders?

[1] In the instrumentation below, (R) denotes an item that is reverse coded.

- Which one do you think is more important for a child to have: obedience or self-reliance?
- Which one do you think is more important for a child to have: curiosity or good manners?
- Which one do you think is more important for a child to have: being considerate or well behaved?

Diversity in Dispositions

In addition to political diversity, we were interested in assessing how the degree of dispositional (i.e. non-political attitudes) differences impacted group decision-making:

1. Need for cognition (Cacioppo, Petty and Feng Kao, 1984; Rathbun, Kertzer and Paradis, 2017)
    - I would prefer complex to simple problems (R)
    - I only think as hard as I have to (R)
    - The idea of relying on thought to make my way to the top appeals to me
    - I really enjoy a task that involves coming up with new solutions to problems
    - Learning new ways to think doesn't excite me very much (R)
    - I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought
    - I feel relief rather than satisfaction after completing a task that required a lot of mental effort (R)
    - It's enough for me that something gets the job done; I don't care how or why it works (R)
2. Big five personality traits (Soto and John, 2017)
    - Extraversion (Tends to be quiet (R); Is dominant, acts as a leader; Is full of energy)
    - Agreeableness (Is compassionate, has a soft heart; Is sometimes rude to others (R); Assumes the best about people)
    - Conscientiousness (Tends to be disorganized (R); Has difficulty getting started on tasks (R); Is reliable, can always be counted on)
    - Neuroticism (Worries a lot; Tends to feel depressed, blue; Is emotionally stable, not easily upset (R))

- Openness to Experience (Is fascinated by art, music, or literature; Has little interest in abstract ideas (R); Is original, comes up with new ideas)

3. Trait aggression (Buss and Perry, 1992*a*)
    - Given enough provocation, I may hit another person
    - I have threatened people I know
    - I often find myself disagreeing with people
    - My friends say that I'm somewhat argumentative
    - Sometimes I fly off the handle for no good reason
    - I have trouble controlling my temper
    - At times I feel I have gotten a raw deal out of life
    - I wonder why sometimes I feel so bitter about things
4. Risk orientations (Ehrlich and Maestas, 2010)
    - In general, people often have to take risks when making financial, career, or other life decisions. Overall, how would you place yourself on the following 7-point scale? [extremely comfortable taking risks $\rightarrow$ extremely uncomfortable taking risks]

Diversity in Demographics

Examining the descriptive diversity of the decision unit, we examined the following features, once again calculating the variance within the group on the following measures:

1. Gender
2. Age
3. Education
4. Income
5. Race (white/non-white)
6. Religion (Christian/non-Christian)

Diversity of Experience

Finally, to assess diversity of experience, we asked about a variety of related measures of experience – in the workforce, in the government, as leaders/supervisors, and as part of a small group team. The first two of these items are derived from Renshon (2017).

1. The maximum number of employees supervised
2. The number of years a person has worked in their career
3. Whether an individual has ever worked for the US government
4. How frequently the individual has had to work in small groups in their career

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.[2] 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

[2]On the virtue of experimental methods in political science more generally, see McDermott (2002).

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. All respondents completed the three modules in the same order (perspective taking, intentionality bias, and then reactive devaluation). As Table 2.2 shows, we find no evidence of treatment spillover effects across the experimental modules.

Table 2.2: No evidence of treatment spillover effects

		Outcome measure		
Group condition	ATE	Risky choice	Intentionality	Support for proposal
Individual	Loss frame	**0.329**	-0.005	0.007
	Fatalities		**0.097**	0.026
	China authored			-0.019
Horizontal	Loss frame	**0.398**	0.020	-0.005
	Fatalities		**0.099**	-0.045
	China authored			**-0.063**
Hierarchical	Loss frame	**0.506**	0.029	0.029
	Fatalities		**0.092**	0.039
	China authored			0.013

Note: ATEs statistically significant at $p < 0.05$ denoted in **bold**

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.

As Appendix §1 shows, the dependent variables were always administered at the individual-level for each respondent, but analyzed in a different manner for each experimental condition, two of which are relatively straightforward (in the individual condition, we record each respondents' choice; in the hierarchical condition, we record the leader's choice). In the horizontal condition, since there isn't a single leader, we ask each group member for their response (either what the group decided, or what respondents themselves decided, based upon whether respondents indicated the group was unanimous or not), which means that our analysis requires some sort of aggregation method or decision rule to aggregate the group members' expressed preferences into a single group decision. While there are any number of potential aggregation method we could use, in the paper and supplementary analysis we consider three different approaches:

- Median voter rule: the group decision is that of the median voter within the group. This is the decision-rule used for our main analysis in the paper.
- Majority vote rule: the group decision is that which obtains the largest number of votes within the group. Groups that do not have a majority vote are excluded from the analysis.
- Unanimity rule: the group decision is unanimously expressed by the group members. Groups that fail to come to unanimity are excluded from the analysis.

For our main analysis in the paper we utilize the median voter rule when analyzing the results from the horizontal condition, not only because it is relatively intuitive, but also because it also doesn't require us to jettison observations and sacrifice statistical power.

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

Table 2.3: Prospect theory experiment

	Individual vs Horizontal				Individual vs Hierarchical		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Loss frame	0.329***	0.323***	0.324***	0.324***	0.329***	0.322***	0.324***
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
Horizontal condition	−0.018	−0.017	−0.058	−0.020			
	(0.040)	(0.041)	(0.053)	(0.041)			
Loss x Horizontal	0.068	0.079	0.198***	0.098*			
	(0.056)	(0.056)	(0.070)	(0.057)			
Hierarchical condition					−0.117***	−0.120***	−0.120***
					(0.042)	(0.043)	(0.043)
Loss x Hierarchical					0.176***	0.173***	0.182***
					(0.057)	(0.058)	(0.058)
Constant	0.438***	0.483***	0.483***	0.487***	0.438***	0.573***	0.489***
	(0.023)	(0.159)	(0.162)	(0.160)	(0.023)	(0.140)	(0.158)
Controls	No	Yes	Yes	Yes	No	Yes	Yes
Horizontal aggregation method	Median voter	Median voter	Unanimity rule	Majority rule			
Hierarchical controls						Leader	Group
N	1,164	1,164	979	1,146	1,127	1,127	1,127
Adjusted R^2	0.133	0.136	0.155	0.140	0.160	0.161	0.162

*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: Intentionality bias experiment

	Individual vs Horizontal				Individual vs Hierarchical		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fatalities	0.097***	0.103***	0.104***	0.103***	0.097***	0.101***	0.102***
	(0.017)	(0.016)	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)
Horizontal condition	0.017	0.014	0.024	0.014			
	(0.022)	(0.022)	(0.035)	(0.022)			
Fatalities x Horizontal	0.003	−0.003	0.027	−0.003			
	(0.031)	(0.030)	(0.050)	(0.030)			
Hierarchical					0.012	0.017	0.007
					(0.021)	(0.021)	(0.021)
Fatalities x Hierarchical					−0.005	−0.006	−0.006
					(0.030)	(0.029)	(0.029)
Constant	0.592***	0.547***	0.551***	0.547***	0.592***	0.538***	0.568***
	(0.012)	(0.079)	(0.083)	(0.079)	(0.012)	(0.071)	(0.080)
Controls	No	Yes	Yes	Yes	No	Yes	Yes
Horizontal aggregation method	Median voter	Median voter	Unanimity rule	Majority rule			
Hierarchical controls						Leader	Group
N	1,044	1,044	838	1,044	1,113	1,113	1,113
Adjusted R^2	0.042	0.110	0.117	0.110	0.037	0.100	0.095

*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 2.5: Reactive devaluation experiment

	Individual vs Horizontal				Individual vs Hierarchical		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
China authored	−0.019	−0.019	−0.019	−0.020	−0.019	−0.021	−0.019
	(0.017)	(0.017)	(0.017)	(0.017)	(0.018)	(0.017)	(0.017)
Horizontal condition	0.064***	0.059**	0.129***	0.067***			
	(0.023)	(0.023)	(0.031)	(0.025)			
China x Horizontal	−0.045	−0.049	−0.084*	−0.049			
	(0.034)	(0.033)	(0.044)	(0.037)			
Hierarchical					−0.008	−0.019	−0.020
					(0.022)	(0.022)	(0.022)
China x Hierarchical					0.032	0.034	0.035
					(0.031)	(0.031)	(0.031)
Constant	0.625***	0.577***	0.557***	0.570***	0.625***	0.470***	0.592***
	(0.013)	(0.084)	(0.084)	(0.085)	(0.013)	(0.075)	(0.084)
Controls	No	Yes	Yes	Yes	No	Yes	Yes
Horizontal aggregation method	Median voter	Median voter	Unanimity rule	Majority rule			
Hierarchical controls						Leader	Group
N	988	988	852	934	1,075	1,075	1,075
Adjusted R^2	0.010	0.057	0.082	0.063	−0.001	0.041	0.042

*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.

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*b*), risk attitudes (Kertzer, 2017), and the "big 5" personality traits (McCrae and Costa, 1987)). Crucially, as the results in Tables 2.3-2.5 show, our results hold both with and without including these controls.

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 2.3-2.5.

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.[3]

[3]There is one additional exclusion criterion in the studies, which we preserve in the analyses below: we remove any group from

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 2.6-2.8, find strikingly similar patterns as the results depicted in the main text (reproduced here in models 4 and 10).

- Second, we then conduct an "extreme bounds" analysis, similar in spirit to Manski bounds (Manski, 1990).[4] This analysis, shown in models 2-3, 6-7 and 12-13 of Tables 2.6-2.8, 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.[5] 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

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.

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

[5] Our 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.

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 2.6-2.8). For hierarchical groups with missing leaders, we similarly replace the missing leader votes with these predicted votes (model 14 in Tables 2.6-2.8). 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 9 in Tables 2.6-2.8).[6] 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.

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 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

[6] In the hierarchical condition, the influencer case is the same as the antisocial case, since the leader’s decision is automatically pivotal by design. Note that there is also a third potential case, where our missing respondents are highly persuadable, and side with a majority of their group. This case is important for our purposes, since for the horizontal condition, this would be equivalent to the results from the main text, since the median vote of each group would remain the same. For the hierarchical condition, this would imply leaders who defer to their advisers. When we estimate the results of this third case in the hierarchical condition, we once again replicate our findings from the main text.

Table 2.6: Prospect theory sensitivity analysis

	Individual			Horizontal			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Loss Trt	0.329***	0.329***	0.329***	0.398***	0.350***	0.348***	0.347***
	(0.033)	(0.033)	(0.033)	(0.043)	(0.032)	(0.032)	(0.032)
Intercept	0.438***	0.438***	0.438***	0.421***	0.435***	0.431***	0.441***
	(0.024)	(0.024)	(0.024)	(0.031)	(0.023)	(0.023)	(0.023)
Model	Base	Extreme Min	Extreme Max	Base	Complete Obs	Extreme Min	Extreme Max
N	760	760	760	404	735	736	736
Adjusted R^2	0.112	0.112	0.112	0.175	0.138	0.136	0.136

	Horizontal		Hierarchical				
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Loss Trt	0.355***	0.355***	0.506***	0.439***	0.435***	0.432***	0.440***
	(0.032)	(0.032)	(0.044)	(0.036)	(0.036)	(0.036)	(0.036)
Intercept	0.432***	0.432***	0.321***	0.366***	0.361***	0.376***	0.368***
	(0.023)	(0.023)	(0.033)	(0.027)	(0.027)	(0.027)	(0.027)
Model	Neural Net Antisocial	Neural Net Influencer	Base	Complete Obs	Extreme Min	Extreme Max	Neural net
N	736	736	367	581	589	589	589
Adjusted R^2	0.142	0.142	0.261	0.199	0.194	0.194	0.200

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table 2.7: Intentionality bias sensitivity analysis

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

	Horizontal		Hierarchical				
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Fatal Trt	0.082***	0.077***	0.092***	0.080***	0.050*	0.072***	0.085***
	(0.015)	(0.014)	(0.025)	(0.023)	(0.028)	(0.022)	(0.018)
Intercept	0.615***	0.618***	0.604***	0.610***	0.498***	0.682***	0.607***
	(0.010)	(0.010)	(0.018)	(0.016)	(0.020)	(0.015)	(0.013)
Model	Neural Net Antisocial	Neural Net Influencer	Base	Complete Obs	Extreme Min	Extreme Max	Neural Net
N	722	722	365	474	589	589	589
Adjusted R^2	0.040	0.042	0.032	0.023	0.004	0.017	0.033

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

Table 2.8: Reactive devaluation sensitivity analysis

	Individual			Horizontal			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
China Trt	−0.019	−0.021	−0.016	−0.063**	−0.089***	−0.087***	−0.045**
	(0.017)	(0.018)	(0.017)	(0.030)	(0.022)	(0.027)	(0.019)
Intercept	0.625***	0.603***	0.637***	0.687***	0.700***	0.483***	0.797***
	(0.012)	(0.014)	(0.013)	(0.020)	(0.015)	(0.019)	(0.013)
Model	Base	Extreme Min	Extreme Max	Base	Complete Obs	Extreme Min	Extreme Max
N	732	760	760	258	540	706	706
Adjusted R^2	0.0003	0.0005	−0.0002	0.013	0.028	0.013	0.007

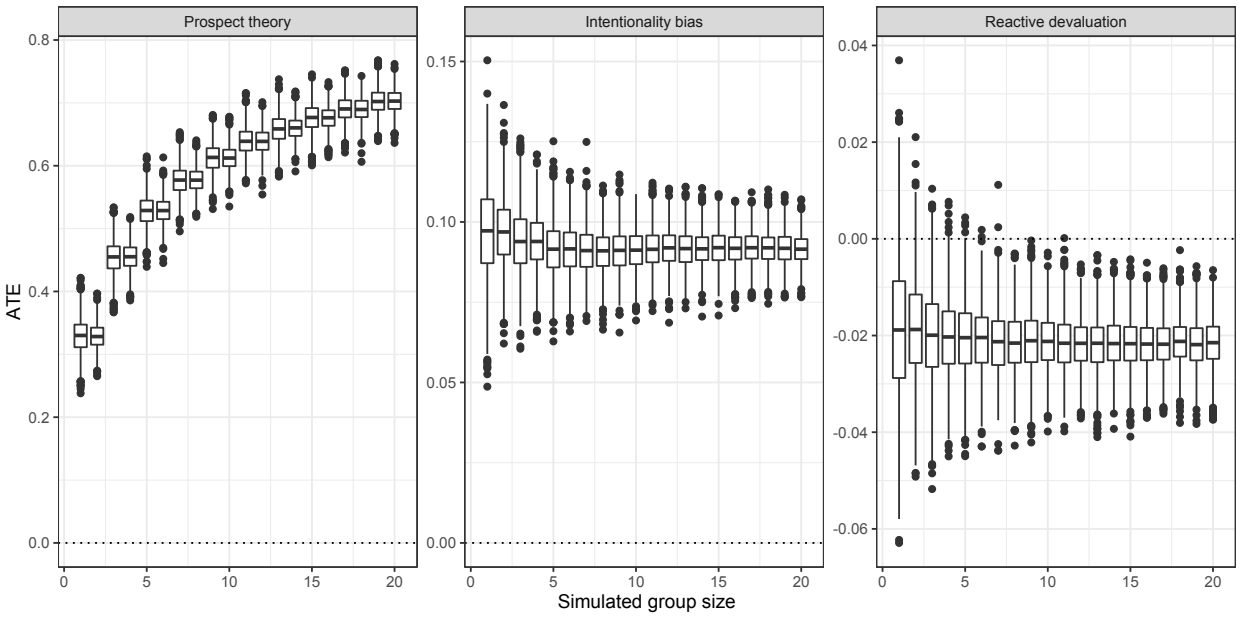
	Horizontal		Hierarchical				
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
China Trt	−0.065***	−0.044***	0.013	−0.004	0.005	−0.008	−0.004
	(0.015)	(0.013)	(0.028)	(0.024)	(0.029)	(0.023)	(0.018)
Intercept	0.678***	0.664***	0.617***	0.620***	0.452***	0.723***	0.620***
	(0.011)	(0.009)	(0.019)	(0.017)	(0.020)	(0.015)	(0.012)
Model	Neural Net Antisocial	Neural Net Influencer	Base	Complete Obs	Extreme Min	Extreme Max	Neural Net
N	706	706	343	433	589	589	589
Adjusted R^2	0.024	0.014	−0.002	−0.002	−0.002	−0.002	−0.002

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

Narang (2017).

- Randomly assign treatment conditions to 1000 groups (e.g. `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.

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.

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 *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).

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 — can also appear 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 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, 76% of group members participate more than once in the deliberation in the prospect theory experiment, 73% do so in the intentionality bias experiment, and 73% 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 74% 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 lengthier discussions; 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 (operationalized as the total number of times group members participated in a given discussion, divided by the number of group members). We then interact this group participation rate with the study-level treatments, and present the results in Table 4.9, 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 4.9 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 ef-

Table 4.9: Little evidence of heterogeneous effects by participation

	Prospect theory			Intentionality bias			Reactive devaluation		
	Horiz.	Hier.	Indiv.	Horiz.	Hier.	Indiv.	Horiz.	Hier.	Indiv.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.420***	0.644***	0.336***	0.058	0.117**	0.150***	−0.038	−0.017	−0.054*
	(0.095)	(0.113)	(0.055)	(0.057)	(0.056)	(0.029)	(0.063)	(0.057)	(0.030)
Group participation	−0.008	0.010		−0.003	0.003		−0.006	−0.014	
	(0.022)	(0.023)		(0.010)	(0.008)		(0.011)	(0.010)	
Trt x Group participation	−0.007	−0.043		0.014	−0.008		−0.007	0.011	
	(0.029)	(0.032)		(0.017)	(0.016)		(0.016)	(0.014)	
Individual participation			0.001			−0.0001			−0.002***
			(0.002)			(0.001)			(0.001)
Trt x Indiv participation			−0.0004			−0.002**			0.001
			(0.003)			(0.001)			(0.001)
Constant	0.444***	0.287***	0.424***	0.618***	0.590***	0.594***	0.706***	0.658***	0.679***
	(0.070)	(0.083)	(0.042)	(0.036)	(0.032)	(0.021)	(0.042)	(0.040)	(0.023)
N	404	367	760	296	358	748	258	333	732
Adjusted R^2	0.173	0.262	0.110	0.041	0.027	0.055	0.011	−0.001	0.011

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

fects 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 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.

5 Ethical considerations

Political scientists have become increasingly interested in ethics in experimental research. Our experimental design decisions were guided by ethical considerations. First, due to anonymity concerns we chose to structure the interactions so that they were text-based rather than video. Second, we ensured that the experimental vignettes did not include any sensitive topics that might cause undue harm to our participants. We also were careful to disclose to respondents that they would be engaged in a discussion with other individuals, providing an opportunity for participants to choose to abandon the survey if this was uncomfortable. Lastly, the studies were approved by Institutional Review Boards (IRBs) at three research universities and at the funder institutions.

References

Altemeyer, Bob. 1981. *Right-wing Authoritarianism*. Winnipeg: University of Manitoba Press.

Becker, Joshua, Douglas Guilbeault and Ned Smith. 2018. "Against Voting? The Crowd Classification Problem." Working paper.

Buss, Arnold H and Mark Perry. 1992*a*. "The aggression questionnaire." *Journal of personality and social psychology* 63(3):452.

Buss, Arnold H and Mark Perry. 1992*b*. "The aggression questionnaire." *Journal of Personality and Social Psychology* 63(3):452–459.

Cacioppo, John T. and Richard E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42(1):116–131.

Cacioppo, John T, Richard E Petty and Chuan Feng Kao. 1984. "The efficient assessment of need for cognition." *Journal of personality assessment* 48(3):306–307.

Ehrlich, Sean and Cherie Maestas. 2010. "Risk Orientation, Risk Exposure, and Policy Opinions: The Case of Free Trade." *Political Psychology* 31(5):657–684.

Feldman, Stanley and Karen Stenner. 1997. "Perceived threat and authoritarianism." *Political Psychology* 18(4):741–770.

Herrmann, Richard K, Philip E Tetlock and Penny S Visser. 1999. "Mass public decisions on go to war: A cognitive-interactionist framework." *American Political Science Review* 93(3):553–573.

Ho, Arnold K, Jim Sidanius, Nour Kteily, Jennifer Sheehy-Skeffington, Felicia Pratto, Kristin E Henkel, Rob Foels and Andrew L Stewart. 2015. "The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO7 scale." *Journal of Personality and Social Psychology* 109(6):1003.

Jost, Tyler. 2021. "Bureaucracy and War: The Institutional Origins of Miscalculation in International Politics." Book manuscript.

Kertzer, Joshua D. 2017. "Resolve, Time, and Risk." *International Organization* 71(S1):S109–S136.

Kertzer, Joshua D., Kathleen E. Powers, Brian C. Rathbun and Ravi Iyer. 2014. "Moral Support: How Moral Values Shape Foreign Policy Attitudes." *Journal of Politics* 76(3):825–840.

Kuhn, Max and Kjell Johnson. 2013. *Applied predictive modeling*. New York, NY: Springer.

LeVeck, Brad L and Neil Narang. 2017. "The democratic peace and the wisdom of crowds." *International Studies Quarterly* 61(4):867–880.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review* 80(2):319–323.

Margalit, Yotam. 2021. "How Markets Shape Values and Political Preferences: A Field Experiment." *American Journal of Political Science* 65(2):473–492.

McCrae, Robert R. and Jr. Costa, Paul T. 1987. "Validation of the Five-Factor Model of Personality Across Instruments and Observers." *Journal of Personality and Social Psychology* 52(1):81–90.

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5:31–61.

Molnar, Andras. 2019. "SMARTRIQS: A Simple Method Allowing Real-Time Respondent Interaction in Qualtrics Surveys." *Journal of Behavioral and Experimental Finance* 22:161–169.

Pratto, Felicia, Jim Sidanius, Lisa M. Stallworth and Bertram F. Malle. 1994. "Social Dominance Orientation: A Personality Variable Predicting Social and Political Attitudes." *Journal of Personality and Social Psychology* 67(4):741–763.

Rathbun, Brian C, Joshua D Kertzer and Mark Paradis. 2017. "Homo diplomaticus: Mixed-method evidence of variation in strategic rationality." *International Organization* 71(S1):S33–S60.

Renshon, Jonathan. 2017. *Fighting for status: Hierarchy and conflict in world politics*. Princeton, NJ: Princeton University Press.

Soto, Christopher J and Oliver P John. 2017. "The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power." *Journal of personality and social psychology* 113(1):117.