What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance At Geopolitical Forecasting[1]

Michael Horowitz, Brandon M. Stewart, Dustin Tingley,
Michael Bishop, Laura Resnick, Margaret Roberts, Welton Chang,
Barbara Mellers, and Phil Tetlock[2] [3]

June 12, 2018

[1] Supplementary material for this article is available in the appendix in the online edition. Replication files are available in the JOP Data Archive on Dataverse. The study was conducted in compliance with relevant laws was reviewed by the appropriate institutional research ethics committee. This research was supported by Research Contract D11PC20061 to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity via the Department of Interior National Business Center. The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government. Research reported in this publication was also supported by The Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number P2CHD047879. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[2] Send comments to: dtingley@gov.harvard.edu.

[3] Michael C. Horowitz is Professor of Political Science at the University of Pennsylvania, Philadelphia, PA 19104; Brandon M. Stewart is Assistant Professor of Sociology at Princeton University, Princeton, NJ 08544; Dustin Tingley is Professor of Government at Harvard University, Cambridge, MA 02421; Welton Chang is a psychologist at the Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723; Michael Bishop is Senior Data Scientist, Shopify, Ottawa, Canada; Laura Resnick is PhD candidate at Columbia University in New York, NY 10025; Margaret Roberts is Associate Professor of Political Science at the University of California at San Diego, La Jolla, CA 92093; Barbara Mellers is the I. George Heyman University Professor at the University of Pennsylvania, Philadelpha, PA 19104; Philip Tetlock is the Annenberg University Professor at the University of Pennsylvania, Philadelphia, PA 19104.

Abstract

How groups make decisions is one of the most fundamental issues in the study of politics. When do groups—be they countries, administrations, or other organizations—more or less accurately understand the world around them and assess political choices? Some argue that group decision-making processes often fail due to groupthink and the biases in decision-making it induces. Others argue groups, by aggregating knowledge, are better at analyzing the foreign policy world. Yet, there is wide variation in how groups perform at processing political information and making accurate forecasts. To advance knowledge about the intersection of politics and group decision-making, this paper draws on evidence from a multi-year geopolitical forecasting tournament with thousands of participants sponsored by the United States government. We find that teams outperformed individuals in making accurate geopolitical predictions, with regression discontinuity analysis demonstrating specific effects from teamwork itself. Moreover, using structural topic models to assess conversations among different teams of forecasters, we find evidence that more cooperative teams outperformed less cooperative teams. Teams that more explicitly engaged in probabilistic reasoning also excelled. These results demonstrate that information-sharing through groups can lead to success in group tasks in the national security community; teams can and do accurately assess the geopolitical world under the right conditions. Moreover, by deliberately cultivating reasoning designed to hedge against cognitive biases and ensuring all perspectives are heard, groups can be more accurate at understanding politics.

Keywords: forecasting, foreign policy, groupthink, topic models

Short title: What Makes Foreign Policy Teams Tick

1 Introduction

The role of groups in decision-making is a critical issue for politics. Nearly all decisions made by governments are the work of groups, not single individuals. Even in strong presidential systems such as the United States, the president rarely makes decisions alone; groups decide which issues make it onto the president's agenda, groups decide how to present the information to the president, and the core decision process is designed to be carried out by groups. Thus understanding how groups make decisions is a key goal for the study of politics.[1]

Strategies that make groups more effective at gathering information, processing it and accurately comprehending the world around them are especially important in the national security realm (Tetlock, 1999). The failure of groups within the U.S. government to accurately assess the likelihood of nuclear tests by India and Pakistan in 1998, the threat posed by international terrorist organizations prior to 9/11, or the state of Iraq's WMD programs in both 1991 and 2003 stand out as some of the most significant intelligence and policy failures of the last several decades. These failures occurred despite the work of teams composed of smart, dedicated individuals who had access to a large amount of information about the world and resources at their disposal (Jervis, 2006). Why then did they fail so spectacularly to understand and decisively act on important geopolitical happenings? One potential explanation for these analytic failures is groupthink, or the rush to conformity of opinion and premature cutoff of debate due to social pressure. Decision making bodies that are unable to engage in effective deliberative thinking are more likely to make bad decisions in a variety of scenarios, especially during foreign policy crises (Janis, 1982; 't Hart, 1990). Groupthink can lead to suboptimal choices when it comes to processing information, predicting the future, and making decisions. Alternatively, with greater access to information, one might expect groups and teams to make better choices than individuals. What are the scope conditions that influence whether groups or individuals make better decisions in the national security arena? This

[1] Groups are typically defined in the literature as units comprising more than two individuals. Likewise teams, a kind of group, are similarly numerically composed, although one key distinction is that team members are generally more familiar with one another than group members, although this is not always the case. While we use the terms groups and teams interchangeably in this paper, we do recognize that the two are conceptualized differently in the literature.

is an especially important question given the high-stakes involved.

To develop a more theoretically and empirically grounded understanding of group and team decision-making within a political context, this paper presents evidence from a geopolitical and economic forecasting tournament with thousands of participants sponsored by the United States government. It yields data on what few studies have before: large scale experiments on real-world forecasting in international affairs using non-student populations. Participants entered predictions about potential geopolitical and economic events, such as whether North Korea would test a nuclear device by a certain date or whether Greece would leave the Eurozone by a certain date. As part of the tournament, participants were randomly selected into team and individual conditions, allowing for a controlled test of the relative effectiveness of teams versus individuals at forecasting geopolitical outcomes. In addition, both teams and individuals were encouraged to explain the reasoning behind their predictions. By evaluating both the reasoning behind the forecasts and the forecasts themselves, we can evaluate the accuracy of teams versus individuals, as well as the conditions under which teams are more likely to succeed or fail. Essentially, the design allows us to identify the situations in which group behaviors such as groupthink, are more likely versus those situations and conditions that set groups up to succeed. This approach makes a significant contribution in part because while group behavior has been subjected to steady investigatory attention, much of the research on group decision-making has been non-experimental. And existing experimental work has often had narrow samples and short timeframes.

Our study uses a purposeful design to advance knowledge: a large-scale randomized-controlled experiment employing a task that resembles, at least in some ways, what national-security personnel at the working level face when they work to put together recommendations for high level decision-makers. This provides the most externally valid test of previous results on the relationship between group behavior and national security decision-making to date, though it does have limits due to the experimental design.[2] Moreover, aside from the macro-theoretical implications, the results have important implications for students of intelligence analysis and political forecasting.

The paper proceeds as follows. Section 2 situates our work in the literature on collective decision-making. Section 3 describes the data gathering project in greater detail and puts

[2] For more on other experimental approaches to international relations, see Mintz *et al.* (2011).

forth our hypotheses. Sections 4 and 5 present the empirical results, showing that not only do teams outperform individuals, but teams featuring broader and deeper engagement are less prone to groupthink-like biases when it comes to geopolitical forecasting. In these sections, a novel application of machine learning methods to the textual data generated by participants allows us to explain how and why some groups succeed while others do not. Section 6 concludes by summarizing our contributions and highlighting areas for future work.

2 Decision-making in International Relations

Both decision-making and forecasting are critical topics in the study of politics. Countries and leaders that make better decisions and forecasts are more likely to succeed in advancing national interests, whether the issue is setting economic policy, designing a military strategy, or deciding whether to sign a free trade deal. Throughout governments, even at very high levels, group processes dominate as the mechanism by which governments make such decisions. For example, in the United States government, important foreign policy decisions generally go through multiple levels of group discussions within the Defense Department, State Department, National Security Council, and elsewhere, as part of what is called the interagency process, before they reach the president. Allison's foundational work on the Cuban Missile Crisis focuses, in part, on this group process and how it shaped US behavior (Allison, 1969). Even in countries with very small selectorates (De Mesquita and Smith, 2005), leaders generally make decisions about important topics such as war and peace within groups.

So, how do groups make decisions? For almost two generations, psychologists have studied variation in group and team decision-making. Beginning with research on excessive conformity (Asch, 1956), research has demonstrated how putting individuals into groups can lead to polarization of opinions (Myers and Lamm, 1976), social loafing and diffusion of responsibility (Darley and Latane, 1968; Karau and Williams, 1993), and typically privilege public information over privately held information even when that information is directly relevant to the task at hand (Stasser and Stewart, 1992). It can also generate questions of social conformity and compliance (Cialdini and Goldstein, 2004). Some of these potential challenges for groups come together under the rubric of groupthink. Groupthink is defined as "[A]mode of thinking that people engage in when they are deeply involved in a cohesive in-group, when the members' strivings for unanimity override their motivation to realistically appraise alternative courses of action" (Janis, 1982, pg. 9). Janis (1982) argues that group pathology in foreign policy decision-making can lead individual members of the group to conform to group norms, rather highlight the diversity of perspectives that should be the strength of groups.[3]

This could have several consequences for group performance in foreign policy and national security settings. First, groups seeking excessive consensus on a decision limit their discussions to only some of the relevant information and thus few courses of action (Janis, 1982; McCauley, 1989; Schulz-Hardt *et al.*, 2000). Second, groups do not adequately examine their favored policy decision in light of non-obvious risks that might not have been considered during initial discussions (Janis, 1982; Janis and Mann, 1977). Third, policy decisions that were initially rejected by the group are never adequately considered (Kahneman and Tversky, 1979). Fourth, groups exhibit selection bias when evaluating new information, ignoring facts that do not support their favored policy proposal (Janis, 1982). Fifth, groups will often fail to discuss contingency plans for what to do if factors arise that might hinder the success of their favored plan (Sunstein and Hastie, 2014; Janis, 1982, see also Janis and Mann 1977, pg. 132). However, the foundational groupthink research used small-n process-tracing approaches to explore the conditions under which group dysfunction could be expected (Janis, 1982; Peterson *et al.*, 1998; Esser, 1998; Tetlock *et al.*, 1992; Schafer and Crichlow, 2013; 't Hart, 1990). This approach makes it difficult to control for the impact of specific antecedents. Experiments on groupthink, for example, have typically involved single-iteration laboratory tasks, without the opportunity to learn from previous mistakes. Furthermore, the experimental tasks were typically undertaken by groups of strangers, a situation that bears little resemblance to the real-world groups that make decisions (a good summary of laboratory experiment results can be found in Esser (1998)).

Yet, there are also reasons to think that groups should be better than individuals at understanding complicated national security questions. There are environments where groups, working together, can produce superior results to those of individuals. In the military context, for example, units with high levels of cohesion generally perform better on the battlefield than those lacking cohesion (Janowitz, 1960). Groups should be a promising environment for decision-making because individuals can bring diverse perspectives to

[3] Additional relevant research includes Sunstein and Hastie (2014); Herrmann (1985); Herrmann and Choi (2008); 't Hart *et al.* (1997); Stern and Sundelius (1997b,a).

the table; the group can then deliberate over the accumulated information, suss out the potential for bias, and arrive at a reasoned conclusion that is better than what a single individual could do (Sunstein and Hastie, 2014). This possibility raises the question of whether different environments might generate different types of practices within groups that make them more likely to be susceptible to groupthink or more likely to embed some of the potentially virtuous practices of groups. Moreover, 't Hart (1990) distinguishes between collective avoidance and collective overoptimism. 't Hart (1990) also notes that group decision-making is useful for things beyond making good decisions– they are used to adjudicate values disputes and to push collective and institutional action.

Teams have been shown to be more creative (Nijstad and De Dreu, 2002; Hoegl and Parboteeah, 2007), take better risks (Rockenbach *et al.*, 2007), and succeed at solving complex problems (Laughlin *et al.*, 2006). Hackman (2002a) points out that good teamwork normally results from proper antecedent conditions, the flipside of Janis focus on the antecedent conditions that lead to groupthink. In addition to being assigned a task that is appropriate for groups to work on, roles such as decision-making authority and structuring incentives such as who benefits and advances, are also important for ensuring harmonious group function (Mathieu *et al.*, 2008). Recently, research on polythink by Mintz and Wayne (2016a,b) highlights that flawed group decision-making processes can emerge even when team members express a plurality of opinions and disagree about the correct policy actions. Note that groups and teams are not necessarily interchangeable, but they are referred to collectively in general in this context because the psychology that motivates them is very similar, and the hypotheses below would apply to groups or teams.

It is also necessary to differentiate between the various units of analysis that are referenced in the research on units larger than pairs. Groups and teams, as opposed to crowds, differ in both group size and in terms of how unit members typically interact. Crowds are typically larger and unit members sometimes do not interact with each other at all (Surowiecki, 2004). Even if there is informational exchange, their ultimate judgments are made independently, thereby inputting pure judgments which when aggregated eliminates error (Larrick *et al.*, 2011). Crowds that are sufficiently expert on the task and are experientially diverse typically outperform individuals (Mannes *et al.*, 2014). Thus, the composition of a crowd has a significant impact on whether the crowd will perform better than the best individual or the average of all of the individuals. This situation is very different from the standard laboratory operationalization and real-world manifestation of what is typically referred to as “small groups” research. While groups come in all shapes and sizes, the hallmark of groups is that they include social interaction of some kind (Hackman and Katz, 2010).

Second, even after the distinction between groups and crowds is established, the literature often uses the terms groups and teams interchangeably (De Dreu and Weingart, 2003; Hackman and Katz, 2010). At least for laboratory tasks, groups are often adhoc and temporary, in comparison to the teams, which are studied in the field and are more stable (Hackman, 2002b; Hackman and Katz, 2010). The key though is that teams and groups are similar in one very important respect: there is an expectation of social interaction (which does not necessarily apply to crowds). This leads to the potential for two motivations to come into conflict: striving to find the truth (in this case, the best judgment or decision) and striving to maintain the group (De Dreu *et al.*, 2008).[4]

While the crowd literature shows that larger units can serve to reduce random error, groups can also (and often do) amplify bias. Kerr *et al.* (1996) reviewed a large body of studies comparing individual and group susceptibility to judgmental biases and found that groups were less susceptible to biases in some cases but in the greater majority of experiments were more susceptible. The ultimate phenomenon is thus: the asymmetrical influence favoring one side, due to a shared conceptual scheme. In some cases, the truth wins and in other cases bias wins (Kerr *et al.*, 1996, 2014). We should expect that teams, especially more stable ones where there is continued expectation of future interaction, would be more susceptible to putting social goals ahead of epistemic ones.

Given this literature, the key question becomes under what conditions groups are more or less likely to succeed at accurately understanding the world around them, particularly in the area of foreign policy. Understanding the overall scope conditions of group decision-making therefore requires not just examining the ability of individuals versus groups to conduct particular tasks, but whether there are conditions that lead to variation in group performance (Hermann, 2012; ’t Hart *et al.*, 1997; Stern and Sundelius, 1997b,a). The next section outlines a novel experiment designed, in part, to test the effectiveness of groups and individuals at forecasting international political events.

[4] The social goal of cohesion is the primary driver of erroneous decision-making in classic groupthink theory (Janis, 1982) and fits with the more recent versions of groupthink that emphasize importance of good group processes for achieving high quality decisions (Schafer and Crichlow, 2010).

3 Project Design and Hypotheses

3.1 Project Overview

This project draws on individual-level forecasts submitted as part of a project funded by the U.S. government, specifically the Intelligence Advanced Research Projects Activity (IARPA), to better understand how to create the most accurate geopolitical forecasts possible.[5]

We use data from 982 individuals. Participants were recruited via e-mail lists, online blogs, and other forums. Participants were required to have a bachelor's degree or higher. There was an attrition rate of 5% over time, so new participants were recruited to ensure that balanced design objectives were reached. On average, 83% of participants were male, 74% were U.S. citizens, and participants had an average age of 40. While the pool was not made up of international politics experts, it did allow the researchers to gather longitudinal experimental data on a non-student population (Mintz *et al.*, 2006). [6]

During each season, IARPA released forecasting questions at regular intervals (generally every few weeks) on geopolitical issues. Forecasting questions were called individual forecasting problems, or IFPs. Examples of questions included: Will NATO invite any new countries to join the Membership Action Plan (MAP) before 1 June 2015? Will Afghanistan sign a Bilateral Security Agreement with the United States before 1 November 2014? For a complete list of questions asked in each season, see the online Appendix.

When new questions were released, participants would log onto a website where they had the option to enter a forecast on each question. For a binary question, such as whether Afghanistan would sign a Bilateral Security Agreement with the U.S, possible forecasts ranged between 0 and 100 (0 = absolutely no, 100 = absolutely yes). Some questions had multiple bins or date ranges where participants would have to enter probabilities in each bin, with the probabilities summing to 100. Importantly, forecasters could log on to the website as often as desired to update their forecasts on all open questions, until that question closed. Any day a forecaster did not log on to update their forecast, their prior

[5] The program was designed as a competition between several teams in industry and at different universities. This article exclusively uses data gathered by Team X.

[6] Based on these demographics, future studies should attempt to create more gender-diverse subject populations, and the results below suggest team processes that include more perspectives can be more effective at times.

forecast on that question carried over to the next day.

Questions closed either when the event posited in the question happened (e.g., Afghanistan signed a Bilateral Security Agreement with the United States), or the question expired without the event occurring. When each question closed, participants received an accuracy score for that question using the Brier scoring rule (Brier, 1950). Brier scores are the sum of the squared deviation between the forecast entered by a participant and the outcome. They range from 0 (perfectly accurate) to .5 (pure chance, such as a coin flip) to 2 (perfectly inaccurate).

As an example, consider the Afghanistan question referenced above. Imagine a participant entered a forecast of 60% for the question of whether Afghanistan would sign a Bilateral Security Agreement with the United States by a certain date on the first day the question was open, and never updated their forecast. The participant would therefore have .60 probability for "yes" and a .40 probability for "no" for each day the question was open. A forecaster gets a score for each day the question is open, based on the final outcome, divided by the number of days the question is open. If Afghanistan did sign a Bilateral Security Agreement with the United States within the time period of the question, therefore, the Brier score for that participant would be $(1 - 0.60)^2 + (0 - 0.40)^2 = 0.32$.

Now suppose that forecaster entered a prediction of 60% the first day the question was open, then updated their prediction to 85% on the 15th day the question was open, and the question closed as "yes" on the 30th day. In that case, the participant would receive 15 days of $(1-0.60)^2+(0-0.40)^2 = 0.32$ and 15 days of $(1-0.85)^2+(0-0.15)^2 = 0.045$, for an overall Brier score on that question of 0.1825. Thus, the faster participants get to the right answer, the better (lower) their Brier score.[7] Participants then received an overall score that was the average of all closed questions, with the top participants arranged, in order, on a leaderboard. Thus, participants could see not only their own scores, but also how their scores compared to the scores of other participants. Participants could see the leaderboard to encourage accountability, encourage competition, and generate transparency.

The experimental design included both individual and group forecasting conditions, providing a robust environment for understanding the influence of group size on forecasting

[7] This is necessary since otherwise, for questions where the potential outcome is not likely to occur, the forecaster could just update their forecast on the last day it closed to the "correct forecast" and receive the same score as someone who got to the right answer weeks earlier.

accuracy. Some participants were randomly assigned into a condition where they made forecasts on their own, while others were randomly assigned into teams of 12-15 members. Individual participants could see a leaderboard of the most accurate forecasters in their experimental condition.[8] Team members communicated through a custom-designed online forum which enabled them to discuss questions and forecast rationales.

Group members entered individual forecasts, with each team receiving a "group" score for each question that was the average of the score of individual members. Group members could also see each other's individual accuracy scores on each question. Groups could also communicate with each other about the research they were doing and their forecasts, so they could deliberate on their predictions together, through a chat room option.

Thus, if an individual on a team disagreed with the way other team members described their forecasts in the online forum, an individual on a team could "defect" from most of the forecasters on their team, enter a different prediction, and then all would be able to judge who was right after the question closed. For participants on teams, the analogue to the leaderboard for participants in the individual experimental condition was a leaderboard featuring the aggregated scores of each team in their experimental condition. By placing some people in teams and having others work alone, repeating interactions over time, and eliciting explanations for their judgments, this is one of the most extensive studies on real-world forecasting in international affairs.[9]

There are differences, of course, between this project and how decisions are actually made in the policy world. For example, many real world teams in the policy and business world make decisions in person, rather than virtually, though the information age has dramatically increased the use of virtual discussions and decision-making. Additionally, due to the structure of the forecasting tournament, all questions had to have explicit end dates, whereas many problems are more indeterminate.

Nevertheless, the evidence about process, and about the relative effectiveness of different kinds of forecasting strategies still has great relevance. The experiment also mimics

[8] Other experimental manipulations included training and in year four, accountability system types.

[9] There was a risk that the use of an intra-team leaderboard could lead to reputation formation and then social loafing. One of the advantages of the experimental design is that we could test the conditions in which this is more or less likely to emerge. The topic models in the results section detail the types of teams where those activities are less likely to occur. As we detail below, a team with that kind of loafing would be less likely to succeed over time.

better than prior research how many in the government talk about problems, such as members of the US intelligence community in diverse locations virtually discussing the strength of a foreign military. The addition of the explicit forecasting task, which often does not happen in the intelligence community at present, is a feature of this design, since, as we describe below, it is the act of forecasting, in part, that helps establish stronger scope conditions for understanding variation in group performance.

Moreover, forecasting problems surrounding important issues in international relations are representative of the uncertainty that many foreign policy professionals face in the real-world but that subjects rarely face in the laboratory. As a result, the opportunity to study group decision making in a setting that not only has difficult tasks but occurs over a long time-span boosts the external validity of any findings.

3.2 Hypotheses

What should explain variation in the ability of groups and individuals to accurately forecast geopolitical events? Group outperformance of individuals is typically seen as task-dependent. Hackman and Katz (2010), in their broad overview of when groups can outperform individuals, point out that compensatory tasks, when the average of the individual inputs is used as the group output, can mitigate the impact of individual biases, resulting in a superior product. Taking the average of the individual inputs also obviates the need to arrive at a forced consensus, thus neutralizing one of the detrimental antecedent conditions of groupthink.

Groups and teams also perform better when they have clear norms for performance (Hackman and OConnor, 2004), which they had in the form of the leaderboard in the experiment. Additionally, several studies demonstrate that virtual teams can perform well because they typically bring to bear a more diverse and knowledgeable group to work on a tough problem (Martins *et al.*, 2004; Powell *et al.*, 2004). It is possible that virtual teams reduce the corrosive effects of social pressures to conformity which enables individuals to speak up and raise their own opinions. Groups were self-organized without assigned leaders who could drive the process. Group members also did not use real names (unless they chose to reveal them, which most did not), instead communicating under usernames. The lack of formal leadership and the ability to operate under a pseudonym reduced the risk that status hierarchies and other related issues could bias group discussion (Sunstein and Hastie, 2014). On the other hand, virtual teams in general do face at least some social pressure, especially when they work together over time. (Woolley *et al.*, 2010). In the national security realm, at the working level, some work increasingly happens through virtual collaboration, though it also occurs in person.

Lastly, groups were incentivized to raise the group's overall accuracy because that score was what ”counted” within the context of the tournament (which was also another benefit to the leaderboard). Group members could see the accuracy score of each groupmate on each question and the overall accuracy of their team (an average of the scores of each member of the team on each question) compared to other teams in their experimental condition. This is the kind of condition that Sunstein and Hastie argue mitigates the effects of groupthink. Teams therefore were incentivized to listen to and follow those team members who had a demonstrated history as most likely to be accurate. By creating status hierarchies based on accuracy, rather than other attributes, forecasting groups were set up in a way to maximize those factors likely to make teams more effective at information sharing and processing. This makes the groups less like decision makers from a foreign policy perespective, but more like working level teams that do much of the work in the foreign policy realm.

From this, we derive the following hypothesis:

Hypothesis 1: *Group forecasters will make more accurate predictions than individual forecasters.*

Even if groups are more accurate than individuals the questions of what sets better performing groups apart from poorer performing groups remains. The question of “why” can help set the scope conditions in which teams are more or less likely to succeed at political forecasting. Research suggests that group and team performance improves when there is equality in the distribution of the conversation (Woolley *et al.*, 2010), rather than following a traditional, vertical hierarchical group process (Sunstein and Hastie, 2014). As previously noted, a tendency towards centralization through the presence of positional and institutional leadership tends to precede groupthink. Additionally, extant literature on group performance shows that decentralized communications and broader group participation leads to improved group performance relative to more centralized and restrictive information flows (Balkundi and Harrison, 2006; Yang and Tang, 2004; Rulke and Galaskiewicz, 2000; Gloor *et al.*, 2008; Leenders *et al.*, 2003). Groups also become more likely to succeed when members cann read the emotions and reactions of others on the team, essentially paying attention to the signals of healthy group interactions that are ever-present in these social settings (even online) (Engel *et al.*, 2014).

Hypothesis 2: *Better performing teams have decentralized conversational norms.*

The research team provided some of its forecasting teams with training in cognitive de-biasing and probability judgments (Mellers *et al.*, 2014). This training included general training to recognize and overcome biases, along with specific encouragement to engage in red teaming and seek out dissenting viewpoints (one of the best practices the literature cited above suggests could lead to more accurate group decisions). In general, the training could be viewed of as a way of priming teams to conduct more metacognition (self-aware thinking about how to think) and complex thinking about how the group itself was making forecasts. Higher levels of self-awareness within groups has been shown to lead to better group performance (Cohen *et al.*, 1996; Kozlowski, 1998; Lord and Emrich, 2001).

Hypothesis 3: *Better performing teams employ metacognition and exhibit higher levels of self-awareness.*

4 Do Teams Matter?

The data analysis below draws on years 2 and 3 of the project. As described above, these years of the project featured individual forecasters as well as forecasters placed in teams. During these years, randomly selected teams and individuals received additional training that focused on cognitive de-biasing as well as how to conduct quantitative probability assessments and study geopolitical issues, the focus of the forecasting tournament. The training programs are described at length elsewhere (Tetlock and Gardner, 2015). In addition to teams and individuals, the experimental design also included a small set of "top performing teams." Top teams selected from the top two percent of forecasters in the preceding year, and all top teams received training in cognitive de-biasing and probability judgments.[10]

To test hypothesis 1 that team performance exceeds individual performance, Section 4.1 analyzes initial forecasting accuracy across experimental conditions. Section 4.2 then uses a regression discontinuity model to test whether the superior performance of top teams established in the previous section can be attributed solely to their composition of "better" individuals, or if something more team-based occurred.

[10] Note that this training, as we explain below, cannot alone explain the variation in the performance of top teams, since other teams also received training. We also explain more about the selection and function of top teams below.

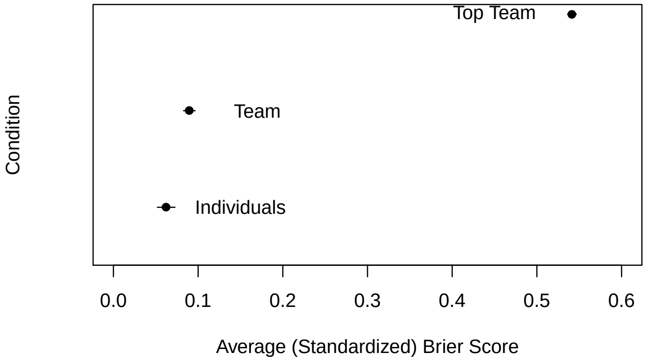

Figure 1: Average (standardized) Brier score by group type. Here and throughout the paper we reverse the standard Brier score so that a higher score indicates greater accuracy.

4.1 Initial Evidence Of Team Performance

We use basic summary statistics to broadly illustrate the performance of teams relative to individuals in the forecasting tournament. A natural starting point is whether groups made better predictions than individuals, on average. Figure 1 plots the average of the standardized Brier score for individuals, teams and top teams[11] with 95% confidence intervals. We reverse the normal Brier scale for presentation purposes, meaning higher scores mean higher levels of accuracy. On average, teams and individuals did the worst, with teams performing slightly better. But top teams had significantly better predictions. Individuals in these teams were able to make substantially superior predictions compared to the other groups and individuals. This parallels findings reported in (Mellers *et al.*, 2015a). On the superiority of teams in the context of the IARPA tournament in general see Mellers *et al.* (2014) and Mellers *et al.* (2015b). As such we find some initial support for hypothesis 1. [12]

[11] Specifically, the score we use is calculated by averaging the Brier scores for all forecasters on a given IFP, and then measuring the standardized deviation from that average for each forecaster on that IFP.

[12] This experimental design is even biased against finding that group decision-making matters, since if explanations force critical thinking that improves accuracy, individuals being primed to provide explanations could increase their accuracy relative to groups where peer pressure presumably means individuals are more likely to explain their reasoning.

4.2 Top Teams Are More Than Top Individuals on the Same Team

The results above do raise the question of whether top teams, which excelled on the metrics above, succeed simply because they are the sum of high-performing parts. Alternatively, is there something in particular about being on a team that improves forecasting accuracy? One way to assess this is to use a regression discontinuity model comparing top team forecasters to nearly identical individuals not on a top team. This allows us to estimate how the exogenous "shock" of joining a top team influences forecasting accuracy for comparable individuals (i.e. those who were close to the qualification threshold). As described above, top teams are composed of forecasters who were the most accurate in their condition the year before they were invited to join a top team. Forecasters who barely miss the cutoff are essentially equivalent to forecasters who barely make the cutoff, because the variation in accuracy between the top forecasters was quite small, substantively; this is why a regression discontinuity design approximates a true experiment in which half of the forecasters near the cutoff were randomly assigned to the treatment, where the treatment is an invitation to join a top team for subsequent years. Promotion to a top team initially occurred in year 2 on the basis of year 1 performance. The top teams construction was designed to see if top performers could replicate that performance in a team setting. Forecasters who made a prediction on at least 45 unique questions and scored in the top 2% of their experimental condition were given an invitation to join a top team in year 2. If the forecaster turned down the invitation, the next most accurate forecaster who met the 45-question threshold was offered their spot. The regression discontinuity analysis is restricted to forecasters who participated in year 1 and year 2, answering at least 45 questions in year 1 and 30 questions in year 2.[13]

Here, we statistically replicate how Team X chose members of the top teams to generate a clear discontinuity we can exploit to test the effect of being placed on a top team. Replicating Team X, we used mean imputed Brier score as the promotion decision criterion, because this was the measure of accuracy that forecasters were incentivized to achieve. The mean imputed Brier score for each individual is a mean across IFPs/question scores

[13] That some possible "top" forecasters turned down the invitation could be seen as biasing the results. Here, however, it provides the means to more effectively test our theory, since we have equivalent forecasters in year 1 placed in separate conditions in year 2.

with an imputed score for the questions a forecaster skipped. In contrast, for assessing year 2 performance, the outcome in our regression, we use the measure that best captures forecasting accuracy: mean standardized Brier score of the questions the individual actually forecast. The standardization is performed at the IFP/question level (i.e., the scores for each question have a mean of zero and standard deviation of one). In both cases, a higher score denotes greater accuracy, because we reverse the normal Brier Score scale for presentation purposes. The promotion decision score criterion (mean imputed Brier score) is a noisier measure of skill than the Brier score, since it includes imputed scores, e.g. participants were assigned the average score for an IFP if they did not enter a forecast themselves. This actually strengthens our regression discontinuity design; it assures us that there truly are forecasters on both sides of the cutoff with equal forecasting skill before some were assigned to top teams.

For ease of interpretation, we center the promotion decision score so those with a score of less than or equal to zero were given an invitation to join a top team. All regression specifications include a centered decision score and a dummy variable indicating whether a forecaster received an invitation to join a top team for year 2. If the dummy variable instead indicated whether the forecaster actually joined a top team in year 2, there would be a threat of selection bias: that the more motivated forecasters accepted the nominations. Our approach, an intent-to-treat analysis, is more conservative, and therefore likely understates the effect of the treatment on the treated. The coefficient on the dummy variable is an unbiased estimate of the causal effect on accuracy of being assigned to a top team and it is large ($\beta = 0.265$, S.E. = 0.035). This means that those selected to participate on top teams relatively improved in subsequent years for reasons related to being on the team, not just because the teams are made up of smart individuals. Graphically, the results can be seen in Figure 2. Here, we fit a loess line to data on each side of the discontinuity. We plot individuals that accepted the top team invitation as open circles. [14]

Sensitivity analyses demonstrate the estimated effect is essentially unchanged even if we include additional measures of accuracy from year 1 and participation in years 1 and 2 (the number of IFPs answered). Furthermore, the result is robust to interacting the promotion decision score and the top team dummy, which represents the possibility that

[14] This also means the solid red dots are forecasters in the individual and team conditions that did not qualify to receive an invitation to join a top team.

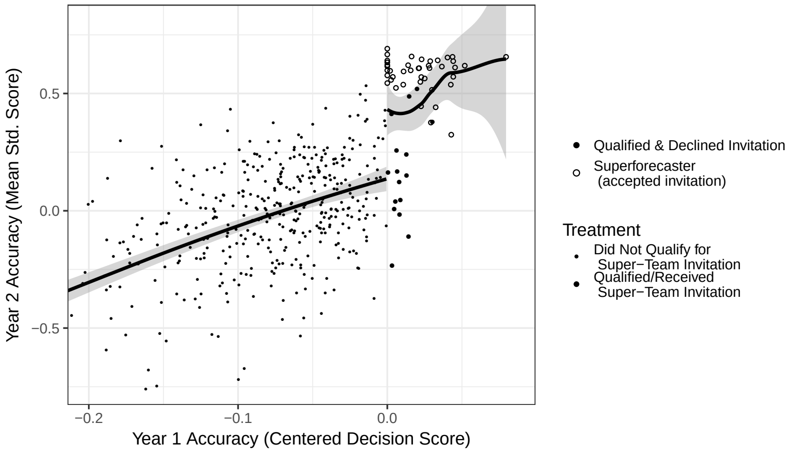

Figure 2: Regression Discontinuity

the slope on the decision score could vary above and below the cut score. These findings further suggest that team dynamics play a significant role in driving top team performance, beyond these teams being an aggregation of talented individuals. This provides especially strong evidence in favor of hypothesis 1.

The regression discontinuity design does not tell us exactly why top teams perform better than counterfactuals near the boundary. Teams could perform well as a result of increased total information available to the group or improved conversations which lead to superior analysis. Indeed the information and analysis effects may not even be conceptually distinguishable in this setting. In the next section, we establish that top teams have distinct patterns of communication which distinguish them from trained and untrained teams.

5 What Kinds of Teams Succeed? Modelling Team Communication

To test hypothesis 2 and hypothesis 3 concerning what explains variation in the ability of groups to forecast, we focus on the content of forecast explanations. In particular, we examine explanations given by individuals in the team conditions. By understanding how different kinds of teams (trained teams, untrained teams, and top teams) use explanations, we can begin unpacking what makes teams more or less effective. We find several patterns in the content of explanations that help to explain top team success.

When making their predictions, participants —whether in the individual or team condition—could also choose to provide an explanation for their forecast. There was a comment box underneath the place where individuals entered their forecasts and participants were encouraged to leave a comment that included an explanation for their forecast. For participants in an individual experimental condition, only the researchers would see those explanations. For participants in a team experimental condition, however, their teammates would be able to see their explanation/comment. These explanations therefore potentially provide useful information to help identify what leads to forecasting accuracy, giving us a way to test hypotheses 2 and 3.

5.1 The Conversational Norms Of Successful Geopolitical Forecasting Groups

An obvious starting point is to ask whether, on average, individuals differ in how extensively they made explanations (i.e., how many comments per IFP) and how intensively (i.e., how long were the comments). Both of these metrics give us a sense of forecaster engagement - since those that explain their predictions are likely more engaged than those that do not. We do this by contrasting behavior by whether a forecaster was on a team or not, whether they were on a team that got training, or not, and whether they were on a top team. Below, we switch from focusing on the extent of engagement to the intensity of engagement, when it occurs.

To calculate the degree of extensive engagement, for each individual we first calculated the total number of explanations made per IFP for which the individual made at least one explanation. Then for each individual we calculated their average number of comments per IFP, averaging over all of the forecasting questions they answered. Thus, for any person we know the average number of explanations they will give for an prediction task.

Figure 3 plots the resulting distribution of this value for each group (individuals, untrained teams, trained teams, and top teams). The x-axis is scaled along a base 10 log for each individual's score because this distribution is heavily skewed. The log transformation reduces the presentational influence of extreme outliers in this distribution. Each group is presented as a different density plot, with the height of the plot giving a relative estimate of how many observations were at the particular value of the x-axis.[15] We observe that both individuals and untrained teams have relatively low levels of average responses per

[15] We use a kernel density function to make the plots.

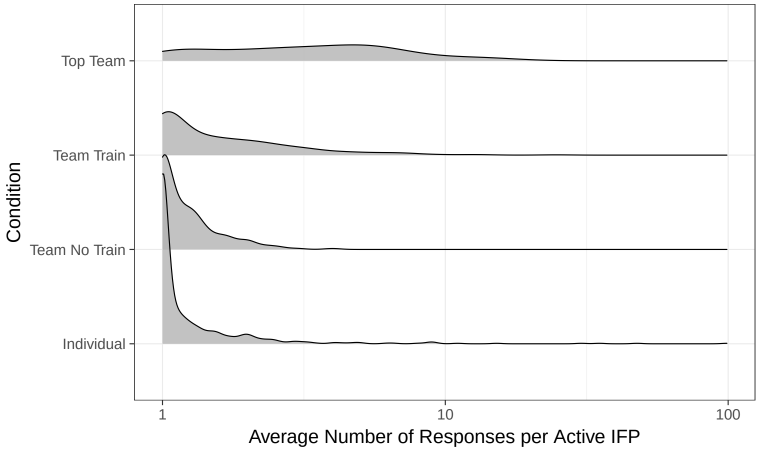

Figure 3: Extensive engagement: number of responses by IFP.

IFP. Trained teams and particularly top teams have considerably higher average responses per IFP.

Next we calculate how intensively individuals engage with explaining their prediction. For each individual we calculated the median length of their first explanation of an IFP. We use the first explanation for a variety of reasons. First, as seen in Figure 3, individuals that were not on a team, or were in untrained teams, rarely made more than one explanation per IFP. Second, we are most interested in individuals providing information and analysis to others on their team. Someone's first explanation is an important first step in doing this. Figure 4 shows the distribution for the four conditions. We see that individuals who are in top teams are clearly engaging in more intensive explanation compared to individuals in other conditions.

Next, we combine Figures 3 and 4 and plot each individual's value of their extensive engagement and intensive engagement in Figure 5. Here we separate out the plots by each of our groups and overlay a contour plot to give a sense of the distribution of data in this space. As expected, we observe that top teams tend to have more individuals who are engaging both more extensively per IFP and more intensively. On the other hand, while people not on teams on occasion would provide multiple explanations per IFP, most did not. Teams with and without training had individuals who provided more lengthy explanations, but these teams do not have individuals who both supplied multiple responses to an IFP and began their engagement with an IFP with a lengthy explanation (which could then be read by other participants on their team).

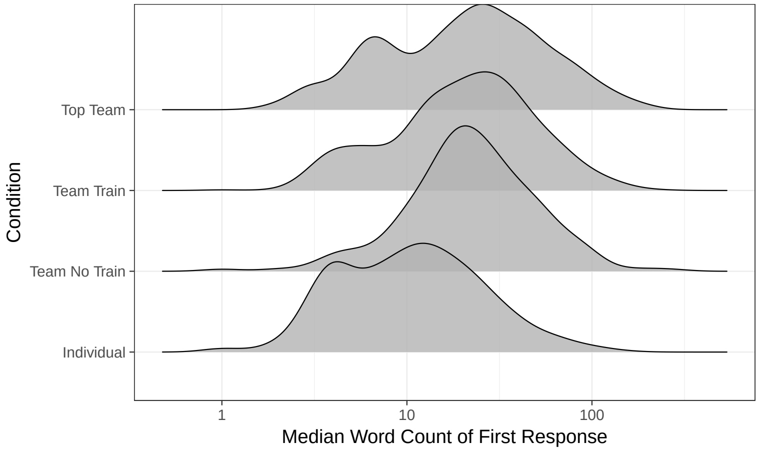

Figure 4: Median number of words used by individuals in their first response to an IFP.

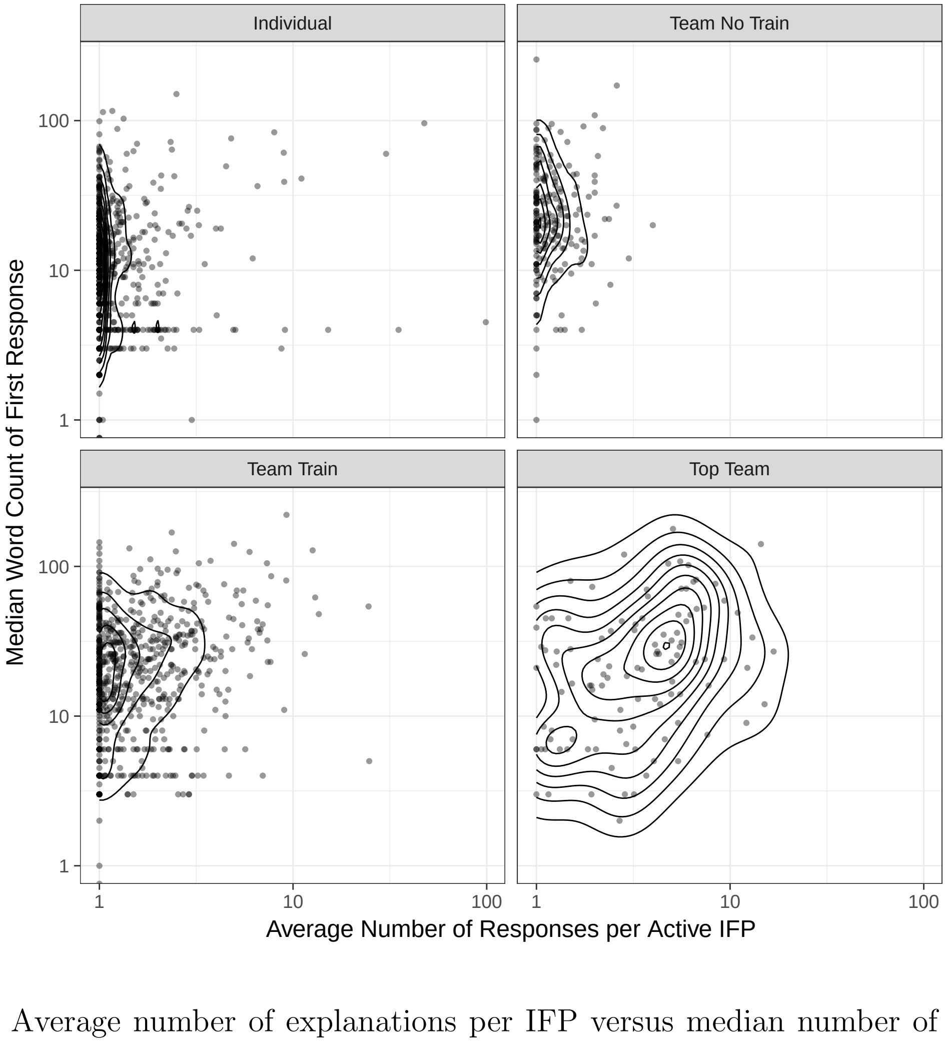

Figure 5: Average number of explanations per IFP versus median number of words for first explanation. X and Y axis are both plotted on log scale.

We also examined other metrics of intensive engagement. Figure 6 plots the fraction of total words in explanations that came after the first response.[16] The plot shows a low proportion of total words coming after the very first explanation from individuals. Teams did better, with more intensive engagement after the first explanation by trained teams and top teams.

Figure 7 investigates the degree to which explanations are generated by a single member of a team or a broader discussion amongst multiple participants. To measure this we calculate for each IFP, in each team, the total number of explanations of the most prolific responder. We then divided this by the average number of responses within the team to that IFP to generate a score for each team/IFP combination. We then plot the distribution of these scores by condition in Figure 7. This shows a distinct pattern illustrating strong effects for one particular type of team - top teams. Prolific posters for top teams posted four times as much as the team average. But for non-top teams, the relative contribution of the most prolific posters was significantly higher. Essentially, in non-top teams, a single person often completely dominates the conversation while top teams featured broader conversations among more team members.[17]

[16] More specifically we calculate this by taking number of words in the first response to an IFP divided by the total number of words in all responses to the IFP. We then subtract that quantity from 1 and take the median for a user across IFPs.

[17] The online appendix looks at whether there are differences in the readability of the explanations. We found no substantive differences across the conditions.

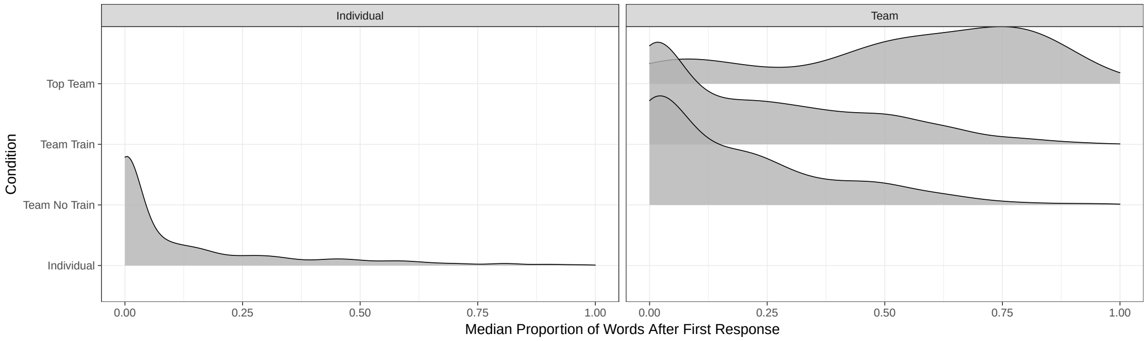

Figure 6: Fraction of total words written after first response.

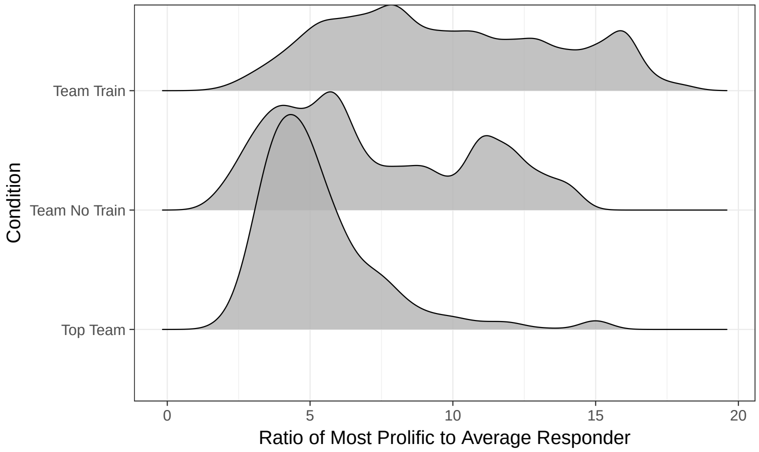

Figure 7: Ratio of responses by the most prolific team member to the team average. Each IFP-team combination is summarized with a score which is the number of posts by the team's most prolific responder to that IFP divided by the team average. The distribution of these scores is then plotted by condition. For example, a score of 5.0 indicates for that IFP-team the most prolific team member responded 5 times as often as the average team member.

That teams, and especially top teams, display a substantial difference in how they engaged with each other provides some evidence for hypothesis 2, because it shows that top teams engaged in both more extensive and intensive engagement, and on average these types of engagement were linked to superior performance through their conversational norms.

Figure 8 then shows that these different conversational norms actually produced greater geopolitical forecasting accuracy. Here, we test hypothesis 2 by evaluating whether that higher degree of engagement on the part of top teams is responsible, in part, for their more accurate forecasting performance. To do this, we first calculate, for each team, on each IFP, the proportion of the team replying (entering a comment on that IFP) and the average of the team's standardized score. We then aggregate over the team-IFP level to the level of the team by taking the median fraction of the team replying (for IFPs in which they participated) and the average score for the team across IFPs. This gives us a sense of the types of team behaviors that lead to better performance overall.

Finally we regressed the measures of accuracy on our team-level extensive engagement score. To allow for any potential non-linear relationships we used a generalized additive

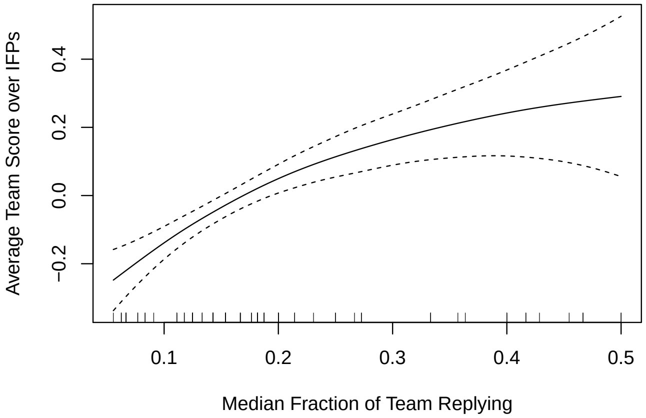

Figure 8: Extensive engagement and performance: Average team score as a function of the median fraction of responses to IFPs by team. Includes controls for team conditions.

model with cubic-regression basis functions, and we plot the 95% confidence intervals.[18] We also control for the effect of condition (team with no training, team with training, and top team).

The results in Figure 8 show an unambiguous positive relationship between extensive engagement within a team and accuracy across IFPs This provides some support for hypothesis 2, which postulated that incorporating the perspective of multiple individuals will improve performance on a team. However, we do see that by the time we reach 40% of a team responding, the relationship flattens out. This flattening is primarily because the teams who have a median response rate beyond 40% are top teams whose positive effect on team accuracy we control for. When not controlling for condition, accuracy continues to increase essentially linearly up to 50% at which point the benefit of additional voices in the conversation declines. This shows, however, that the extremely hierarchical, top-down conversational patterns that often only feature a few voices are less successful, on average, at comprehending and forecasting on important political questions.

5.2 Metacognition and Geopolitical Forecasting

To examine hypothesis 3 we look at whether the use of metacognition, e.g. thinking about thinking, explains the success of some teams, since they are more self-aware in ways that help them discard biases and evaluate the world more accurately, we turn to

[18] Implemented in the `mgcv` package with option `bs="cs"` (Wood, 2011).

text analysis. There are many ways to analyze text, and available tools are constantly evolving. To assess the explanations offered as part of Team X forecasts, we focus on an unsupervised machine learning technique known as the Structural Topic Model, which has been used in a variety of applications in the social sciences (Roberts *et al.*, 2014, 2018). Topic models are a class of models that discover sets of words that tend to occur together. These co-occurrence patterns allow us to estimate distributions over words called "topics" where each document is a distribution of the estimated topics.

Unlike most existing forms of topic models, such as the popular latent Dirichlet allocation model (Blei *et al.*, 2003), the Structural Topic Model allows for information about individual documents to be incorporated into the estimation of topics. This allows the researcher to investigate the presence of relationships between this "meta-data", information about the documents, and topics of interest.[19] In the current application we use information about whether an explanation came from an individual who was on an untrained team, a trained team, or a top team. We also include indicator variables for each IFP in the analysis that help to pick up domain specific-language.

5.2.1 Sample and Preprocessing

Given our interest in what makes teams most effective, we subset our data to focus only on teams and drop individuals not on teams. We also include only an individual's first response to an IFP because, as discussed before, there is considerable variation across conditions in terms of how frequently individuals would post explanations per IFP. This does not mean someone would post an explanation only after having seen posts from other teammates. Indeed, as we discuss below, we frequently saw individuals engaging with explanations posted by other teammates.

We also pre-processed the data in several ways. First, we only included words that appeared in a minimum of 20 documents. This eases estimation of the model by reducing the total number of words that can be associated with topics. In order to capture linguistics patterns that are common across IFPs rather than specific to the content of the questions, we only included words that appeared in explanations at least twice for at least 10 different IFPs. We also conducted standard processing of textual data such as stemming (processing words that reflect the same concept to a single root) and stopword

[19] Importantly, and as discussed at length in Roberts *et al.* (2014), this approach does not force there to be relationships.

(of, the, etc.) removal.[20]

5.2.2 Results

To estimate the STM we need to set the number of topics ahead of time. A larger number of topics permits a more granular view whereas a smaller number of topics produces a broader view of the corpus being analyzed. We estimated a structural topic model in which we set the number of topics at 45. This allows for a relatively granularly view while not overwhelming the analyst. Estimation with similar numbers of topics generally produced similar results.[21] The online appendix provides a summary of the topics recovered by the model.

Here we investigate several of the topics in greater detail. The top row in Figure 9 gives the top words associated with a teammate topic (Topic 1) as well as two additional topics that we refer to as "analysis" topics, Topics 12 and 39. For each topic we present a "word cloud" representation, where larger words were more highly associated with the topic. The interpretation of the teamwork topic is straightforward. Explanations using this topic had people mentioning that they were were following their teammates and learning from them. In doing so, they would explain that they were benefiting from the research done by their teammates. On occasion they would thank specific individuals by name.

The analysis topics pick up on individuals explaining their arguments in more detail, often using the type of logical and probabilistic reasoning tools that previous research suggests lead to better predictions (Mellers *et al.*, 2014). Topic 12 picks up on individuals sharing information. While the words displayed in Figure 9 help to convey this, it is always useful when using topic models to also look at example documents that are heavily associated with a topic.[22] Topic 39 contains a number of words associated with probabilistic reasoning. Earlier research suggests that predictions that do not admit un-

[20] See Grimmer and Stewart (2013) for additional discussion.

[21] We used an initialization based on the spectral method of moments estimator of Arora *et al.* (2012) to bypass mulimodality issues that naturally arise in topic models Roberts *et al.* (2016).

[22] For example, "Wasn't able to find a lot of information on this one but an informal poll site gave me the 65/35 number so I'll go with it until I get more info" and "Following teammates br/ I'm following [name]'s lead here with a 55 but also noting how hard it is to find any news articles on any India Peru meetings scheduled to discuss this or any of news about a possible deal that is recent. All of [name]'s articles that mention an Indian Peru agreement are from last year and I couldn't find anything from this year that said anything new".

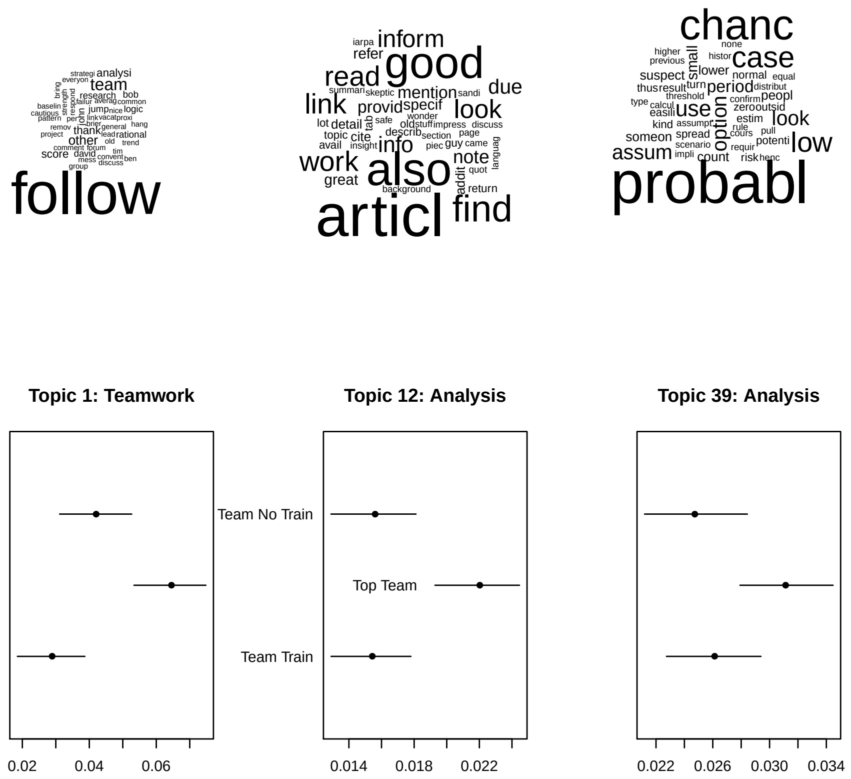

Figure 9: Top words for three topics from 45 topic STM model (top row). Marginal effect of treatment conditions on topic prevalence (bottom row).

certainty one way or the other are likely driven by biases that are unhelpful for ultimate prediction performance.[23]

Interestingly, in the examples quotes individuals are providing analysis but also engaging with teammates. This came in the form of drawing on resources from teammates or in the form of an individual inquiry about the probability predictions of their teammates. This illustrates how the model allows for any explanation to be a mixture of topics. But

[23] An example quote of this topic: “I pulled out my old random walk model If you take an annual std dev of 9 the dev for this year 5 2 3 yr dev is more like 7 then the chance of hitting 105 assuming up/down moves of 1 std is about 14 Chance of hitting 110 is about 2. [link to Google document] If you look at option prices the chance of hitting 105 is also about 10 My calculations are not the most rigorous so pls critique But I think it’s roughly right.”

it also suggests that perhaps teamwork might be particularly effective if it is combined with analysis, illustrating the use of metacognition.[24]

What is the correlation between the explanation an individual gives for their predictions and their forecasting accuracy on a particular forecasting question? This can provide additional evidence to explore hypothesis 3 because it shows what types of teamwork lead to more accurate decision-making. We therefore investigate whether individuals who engage with their teammates and utilize source information and probabilistic reasoning in their analysis perform better on predictions. To measure performance on a prediction task, we use an individual's final prediction score standardize by prediction task. We scale this measure such that higher scores are more accurate. We regress this dependent variable on teamwork topic 1, one of the analysis topics, an interaction between the two, and a set of control variables. In particular, we control for the number of days since the prediction was first posted, a dummy variable for whether the prediction came from year 2 or 3, the overall length of the explanation, and fixed-effects for each prediction task.

Results produced for each of the analysis topics demonstrates how top teams use metacognition to excel, illustrating a key cognitive pathway whereby groups can more effectively forecast on geopolitical issues. Table 1 presents two models that exclusively use the top teams. The first interacts Topic 1 and 12 and the second model interacts Topic 1 and 39. The key result from each of these tables is a positive interaction between the teamwork topic and the analysis topic. This relationship was only present for top teams. One factor that appears to make the team process of top teams much more effective is a simultaneous engagement with one's own analysis as well as the views of others, illustrating the hypothesized metacognition process.

[24] For example, the following entry was highly connected to both topic 1 and 12. "Following teammates. I was planning to abstain from this question but after reading [name] links and recent stuff from gust it appears that the current path is broken because AKP doesn't have quite enough muscle to get it through even to a referendum."

	1	2
Intercept	-0.99^{***}	-0.98^{***}
	(0.07)	(0.07)
daysince	-0.00^{*}	-0.00^{**}
	(0.00)	(0.00)
year	0.50^{***}	0.50^{***}
	(0.04)	(0.04)
length	0.00^{*}	0.00^{*}
	(0.00)	(0.00)
Topic1	-0.05	-0.02
	(0.03)	(0.02)
Topic12	-0.01	
	(0.14)	
Topic39		-0.18^{*}
		(0.10)
Topic1 * Topic12	2.86^{**}	
	(1.15)	
Topic1 * Topic39		2.21^{*}
		(1.18)
Num. obs.	8662	8662
R^2	0.79	0.79
Adj. R^2	0.78	0.78
L.R.	13489.21	13489.99

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 1: OLS models of topics and accuracy

5.3 Summary

Investigations of the patterns and contents of team communication show strong support for hypotheses 2 and 3 that successful teams are decentralized and employ a discourse including metacognition. Successful teams have members who engage more intensively (length of response) and extensively (number of explanations). Particularly distinctive of the highly successful top teams is continued engagement after the first explanation.

Importantly for top teams these discussions are actually team efforts; the most prolific responder typically gives 3-6 times as many explanations as the average member of their team. By contrast, around half of IFPs for other teams (trained and untrained alike) have the most prolific responder contributing 8 or more times the team average. These team dynamics translate directly into accuracy. Teams with a greater fraction of their members responding to IFPs have better forecasting scores even when controlling for top team membership and training.

The results also demonstrate that it is not just the patterns of engagement but the content of that engagement which lead to higher accuracy. For top teams, a combination of discourse on teamwork and analysis is a strong predictor of high forecast accuracy.

This suggests that more than teamwork alone, a key predictor of group success in forecasting on national security issues is the combination of that teamwork with discussion and contribution of new analysis of the problem at hand.

6 Conclusion

Whether groups can accurately assess the world and make good decisions is a vital question in the study of politics. In this paper, we use data from a multi-year geopolitical forecasting effort featuring thousands of forecasters—some working as individuals, some as teams—making hundreds of thousands of predictions. In contrast to the gloomy expectations of some of the literature on group decision-making, including groupthink, we find that forecasting teams far outperformed individuals at accurately predicting diverse geopolitical events, including whether North Korea would test a nuclear weapon by a certain date and who would win elections in countries around the world.

Simple statistical analysis shows that teams outperform individuals, suggesting that there are ways to overcome groupthink. We further illustrate that group success is not simply due to putting high-performing individuals together with a regression discontinuity design that demonstrates how, for forecasters of relatively similar ability, being placed on a team led to more forecasting success.

The results also demonstrate why some groups succeed while others fail. More successful teams were more engaged—with members making more comments per person and with more members of the team commenting. In some ways, this may reflect familiar patterns from politics and business. Hierarchical teams where only a few people speak, dominating the conversation, are, on average, less successful than teams that accept input from a broader representation of team members. Moreover, teams that more effectively employed training on cognitive de-biasing and probability judgments, demonstrated with topic models, were more accurate than those that did not. More accurate groups not only feature individual analysis, though, but genuine teamwork where individuals react to, and update their beliefs, in response to the arguments made by their teammates. This also makes a contribution by showing how reputation formation, rather than leading to social loafing on successful teams, becomes information that teams integrate to continue updating and succeed.[25]

[25] This suggests a linkage to constructivist accounts of how social dynamics influence behavior, a contribution of this paper and a potential avenue for future research.

While our results shed light on the power of teams for forecasting tasks, there are some natural limitations. While we have causally identified evidence that teams and those given training make more accurate predictions, our analysis of the characteristics of successful and unsuccessful teams are necessarily observational. While the evidence is suggestive that teams are engaging in a collective analytic exercise, we hope that future work will further study these mechanisms. We are particularly interested in distinguishing between the effect of introducing new information into the team and improving analysis of the same set of information, however a well-identified study of this distinction would require a fundamentally different study structure. The challenges of operating within a government bureaucracy also mean that there are limits to the application of these results in the real world, since there are some decisions that happen through consensus-forming.

The external validity of the results still do suggest broad applicability in a few dimensions. First, the results about training and conversations illustrate potential methods for training intelligence analysts, foreign service officers, and others to more accurately see the world around them. Second, our results show how team design matters, which could shape how governments make choices about creating working teams. Application of the results of the experiment can be seen today, at least on a small scale, in the existence of ongoing quantitative forecasting efforts, which did not exist before, in the US intelligence community. These efforts seek to complement traditional intelligence analysis with prediction markets and teaming designed to more explicitly build forecasting track records.

Many major government decisions occur through a group process, whether through working-level teams conducting analysis or the results of that analysis being considered by groups of decision-makings. Thus, these results are striking and important. Theoretically, these results suggest new pathways for case studies on effective versus ineffective government decision-making. They also show that, rather than framing questions in terms of whether groups succeed or fail, research should focus on scope conditions. Finally, these results also offer potential lessons for how to improve the ability of groups within the government to understand and anticipate world events, certainly including but not necessarily limited to geopolitical forecasting.

References

Allison, G. T. (1969). Conceptual models and the cuban missile crisis. *The American Political Science Review*, **63**(3), 689–718.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2012). A practical algorithm for topic modeling with provable guarantees. *arXiv preprint arXiv:1212.4777*.

Asch, S. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, **70**(9), 1.

Balkundi, P. and Harrison, D. A. (2006). Ties, leaders, and time in teams: Strong inference about network structures effects on team viability and performance. *Academy of Management Journal*, **49**(1), 49–68.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research*, **3**, 993–1022.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**(1), 1–3.

Cialdini, R. and Goldstein, N. (2004). Social influence: compliance and conformity. *Annual review of Psychology*, **55**, 591–621.

Cohen, M. S., Freeman, J. T., and Wolf, S. (1996). Metarecognition in time-stressed decision making: Recognizing, critiquing, and correcting. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **38**(2), 206–219.

Darley, J. M. and Latane, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology*, **8**(4p1), 377.

De Dreu, C. K. and Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. *Journal of applied Psychology*, **88**(4), 741.

De Dreu, C. K., Nijstad, B. A., and van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, **12**(1), 22–49.

De Mesquita, B. B. and Smith, A. (2005). *The logic of political survival*. MIT press.

Engel, D., W., W. A., Jing, L. X., Chabris, C. F., and Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? theory of mind predicts collective intelligence equally well online and face-to-face. *PloS one*, **9**(12), e115212.

Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational behavior and human decision processes*, **73**(2), 116–141.

Gloor, P. A., Paasivaara, M., Schoder, D., and Willems, P. (2008). Finding collaborative innovation networks through correlating performance with social network structure. *International Journal of Production Research*, **46**(5), 1357–1371.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, **21**(3), 267–297.

Hackman, J. (2002a). Why teams dont work. *Theory and Research on Small Groups*, pages 245–267.

Hackman, J. R. (2002b). *Leading teams: Setting the stage for great performances*. Harvard Business Press.

Hackman, J. R. and Katz, N. (2010). Group behavior and performance. *Handbook of social psychology.*

Hackman, J. R. and OConnor, M. (2004). *What makes for a great analytic team? Individual vs. team approaches to intelligence analysis*. Intelligence Science Board, Office of the Director of Central Intelligence, Washington, DC.

Hermann, C. (2012). *When things go wrong: foreign policy decision making under adverse feedback*. Routledge.

Herrmann, R. (1985). *Perception and behavior in Soviet foreign policy*. University of Pittsburgh Press.

Herrmann, R. and Choi, J. K. (2008). From prediction to learning: Opening expert minds to unfolding history. *International Security*, **31**(4), 132–161.

Hoegl, M. and Parboteeah, K. P. (2007). Creativity in innovative projects: How teamwork matters. *Journal of Engineering and Technology Management*, **24**(1), 148–166.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes*. Houghton Mifflin Boston.

Janis, I. L. and Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment.* Free Press.

Janowitz, M. (1960). *The Professional Soldier*. Free Press.

Jervis, R. (2006). Reports, politics, and intelligence failures: The case of iraq. *Journal of Strategic Studies*, **29**(1), 3–52.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, **47**(2), 263–291.

Karau, S. J. and Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, **65**(4), 681–706.

Kerr, N. L., MacCoun, R. J., and Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological review*, **103**(4), 687.

Kerr, N. L., MacCoun, R. J., and Kramer, G. P. (2014). when are n heads better (or worse) than one?: biased judgment in. *Understanding Group Behavior: Volume 1: Consensual Action By Small Groups; Volume 2: Small Group Processes and Interpersonal Relations*, **1**, 105.

Kozlowski, S. W. (1998). Training and developing adaptive teams: Theory, principles, and research. In J. A. Cannon-Bowers and E. Salas, editors, *Decision making under stress: Implications for training and simulation*, pages 115–153. APA Books.

Larrick, R. P., Mannes, A. E., Soll, J. B., and Krueger, J. (2011). The social psychology of the wisdom of crowds.

Laughlin, P. R., Hatch, E. C., Silver, J. S., and Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: effects of group size. *Journal of Personality and social Psychology*, **90**(4), 644.

Leenders, R. T. A., Van Engelen, J. M., and Kratzer, J. (2003). Virtuality, communication, and new product team creativity: a social network perspective. *Journal of Engineering and Technology Management*, **20**(1), 69–92.

Lord, R. G. and Emrich, C. G. (2001). Thinking outside the box by looking inside the box: Extending the cognitive revolution in leadership research. *The Leadership Quarterly*, **11**(4), 551–579.

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, **107**(2), 276.

Martins, L. L., Gilson, L. L., and Maynard, M. T. (2004). Virtual teams: What do we know and where do we go from here? *Journal of management*, **30**(6), 805–835.

Mathieu, J., Maynard, M. T., Rapp, T., and Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of management*, **34**(3), 410–476.

McCauley, C. (1989). The nature of social influence in groupthink: Compliance and internalization. *Journal of Personality and Social Psychology*, **57**(2), 250.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., *et al.* (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, **25**(5), 1106–1115.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., *et al.* (2015a). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, **10**(3), 267–281.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., and Tetlock, P. (2015b). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, **21**(1), 1.

Mintz, A. and Wayne, C. (2016a). The polythink syndrome and elite group decision-making. *Political Psychology*, **37**(S1), 3–21.

Mintz, A. and Wayne, C. (2016b). *Polythink Syndrome: U.S. Foreign Policy Decisions On 9/11, Afghanistan, Iraq, Iran, Syria, and ISIS*. Stanford University Press.

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

Mintz, A., Yang, Y., and McDermott, R. (2011). Experimental approaches to international relations. *International Studies Quarterly*, **55**(2), 493–511.

Myers, D. G. and Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, **83**(4), 602.

Nijstad, B. A. and De Dreu, C. K. (2002). Creativity and group innovation. *Applied Psychology*, **51**(3), 400–406.

Peterson, R. S., Owens, P. D., Tetlock, P. E., Fan, E. T., and Martorana, P. (1998). Group dynamics in top management teams: Groupthink, vigilance, and alternative models of organizational failure and success. *Organizational behavior and human decision processes*, **73**(2), 272–305.

Powell, A., Piccoli, G., and Ives, B. (2004). Virtual teams: a review of current literature and directions for future research. *ACM Sigmis Database*, **35**(1), 6–36.

Roberts, M., Stewart, B., and Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez, editor, *Computational Social Science: Discovery and Prediction*, pages 51–97. Cambridge University Press, New York.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., and Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2018). stm: R package for structural topic models. *Journal of Statistical Software*.

Rockenbach, B., Sadrieh, A., and Mathauschek, B. (2007). Teams take the better risks. *Journal of Economic Behavior & Organization*, **63**(3), 412–422.

Rulke, D. L. and Galaskiewicz, J. (2000). Distribution of knowledge, group network structure, and group performance. *Management Science*, **46**(5), 612–625.

Schafer, M. and Crichlow, S. (2010). *Groupthink versus high-quality decision making in international relations*. Columbia University Press.

Schafer, M. and Crichlow, S. (2013). *Groupthink versus high-quality decision making in international relations*. Columbia University Press.

Schulz-Hardt, S., Frey, D., Lüthgens, C., and Moscovici, S. (2000). Biased information search in group decision making. *Journal of personality and social psychology*, **78**(4), 655.

Stasser, G. and Stewart, D. (1992). Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, **63**(3), 426.

Stern, E. K. and Sundelius, B. (1997a). Foreign policy-making at the top: Political group dynamics. In P. 't Hart, E. K. Stern, and B. Sundelius, editors, *Beyond Groupthink: Political Group Dynamics and Foreign Policy-making*, pages 3–34. University of Michigan Press.

Stern, E. K. and Sundelius, B. (1997b). Understanding small group decisions in foreign policy: Process diagnosis and research procedure. In P. 't Hart, E. K. Stern, and B. Sundelius, editors, *Beyond Groupthink: Political Group Dynamics and Foreign Policy-making*, pages 123–150. University of Michigan Press.

Sunstein, C. R. and Hastie, R. (2014). *Wiser: Getting beyond groupthink to make groups smarter*. Harvard Business Press.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter and how collective wisdom shapes business, economies, societies, and nations.

't Hart, P. (1990). *Groupthink in government: a study of small groups and policy failure*. Swets and Zeitlinger.

't Hart, P., Stern, E. K., and Sundelius, B., editors (1997). *Beyond Groupthink*. University of Michigan Press.

Tetlock, P. E. (1999). Theory-driven reasoning about plausible pasts and probable futures in world politics: are we prisoners of our preconceptions? *American Journal of Political Science*, pages 335–366.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown.

Tetlock, P. E., Peterson, R. S., McGuire, C., Chang, S.-j., and Feld, P. (1992). Assessing political group dynamics: a test of the groupthink model. *Journal of personality and social psychology*, **63**(3), 403.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(1), 3–36.

Woolley, A. W., Chabris, C. F., Pentland, A., Nashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, **330**(6004), 686–688.

Yang, H.-L. and Tang, J.-H. (2004). Team structure and team performance in is development: a social network perspective. *Information & Management*, **41**(3), 335–349.

7 Affiliations

Michael C. Horowitz is Professor of Political Science at the University of Pennsylvania, Philadelphia, PA 19104; Brandon M. Stewart is Assistant Professor of Sociology at Princeton University, Princeton, NJ 08544; Dustin Tingley is Professor of Government at Harvard University, Cambridge, MA 02421; Welton Chang is a psychologist at the Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723; Michael Bishop is Senior Data Scientist, Shopify, Ottawa, Canada; Laura Resnick is PhD candidate at Columbia University in New York, NY 10025; Margaret Roberts is Associate Professor of Political Science at the University of California at San Diego, La Jolla, CA 92093; Barbara Mellers is the I. George Heyman University Professor at the University of Pennsylvania, Philadelpha, PA 19104; Philip Tetlock is the Annenberg University Professor at the University of Pennsylvania, Philadelphia, PA 19104.