The World Is More Uncertain Than You Think: Assessing and Combating Overconfidence Among 2,000 National Security Officials

Jeffrey A. Friedman
Associate Professor of Government, Dartmouth College

Forthcoming at the *Texas National Security Review*

Abstract. This article analyzes more than 60,000 assessments of uncertainty made by national security officials from more than forty NATO allies and partners. It shows that national security officials are overwhelmingly overconfident and that their judgments are especially prone to false positives. Despite having strong incentives to make accurate assessments of uncertainty, national security officials share biases that are widespread among the general public. These flaws also appear to be tractable: just two minutes of training significantly improved performance. Altogether, these findings demonstrate how national security bureaucracies can leverage insights from the decision sciences to improve cognitive performance at large scales.

Acknowledgements. Special thanks go to the 1,894 national security officials who generously volunteered their time to participate in this study. Rich Andres, Christina Brookes, Mark Bucknam, Jennifer Lerner, Stephen Mariano, Idun Mustulien, and Bryan Pendleton played irreplaceable roles in facilitating the partnerships on which the project depended. Many of the core ideas presented in this article were inspired by Richard Zeckhauser, who also participated in early training sessions. Erik Lin-Greenberg, Nicholas Miller, Caleb Pomeroy, Alberto Simpser, Megan Stewart, Michael Poznansky, and John Wilcox provided thoughtful comments on prior drafts. Freya Jamison, Luca Fagotti, Joowon Kim, and Benjamin Rutan provided research assistance. The research presented in this article was partly conducted while the author was a Visiting Fellow at the Institute for Advanced Study in Toulouse. Funding from the French Agence Nationale de la Recherche (under the Investissement d'Avenir programme, ANR-17-EURE-0010) is gratefully acknowledged.

The World is More Uncertain Than You Think

Uncertainty surrounds virtually every element of international politics. Heads of state confront uncertainty when judging how their counterparts will react to crises.[1] Generals confront uncertainty when evaluating the chances that their strategies will succeed or fail.[2] Intelligence analysts confront uncertainty when they assess other states' capabilities and intentions.[3] Diplomats confront uncertainty when they attempt to discern their negotiating partners' bottom lines.[4] In these circumstances, and many others like them, national security officials must constantly grapple with the fact that they possess imperfect information about the world.

How well do national security officials meet that challenge, and how might that shape international affairs? Scholars answer this question in many different ways. Realists typically argue that national security officials can be expected to assess uncertainty in a rational, unbiased manner.[5] Political psychologists often claim that national security officials are prone to overconfidence, in the sense that they consistently assign too much certainty to their judgments.[6] Overconfidence is widely viewed as a source of instability in international politics, as leaders who exaggerate the chances that their policies will succeed should also be more prone to initiating military disasters.[7] Other scholars, however, argue that national security officials are prone to underconfidence due to professional cultures that discourage taking analytic risks.[8] Underconfidence can undermine national security decision-making, too, particularly by discouraging leaders from exploiting feasible opportunities to advance their country's interests.

It is notoriously difficult to understand which of these problems predominates, and to what extent, in national security decision-making. Part of the problem is that national security is so complex that it is often impossible to say whether any assessment of uncertainty in this domain is "right" or "wrong."[9] For example, if a general claims there is a seventy percent chance they will win a battle, but they lose, it is hard to know whether the general's judgment was flawed or if they simply got unlucky.[10] The standard way to solve that challenge is to evaluate the accuracy of many judgments at once. Thus, if we look at all of the battles where generals predict a seventy percent

---

[1] Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics* 30, no. 2 (1978), pp. 167-214.
[2] Alan Beyerchen, "Clausewitz, Nonlinearity, and the Unpredictability of War," *International Security* 17, no. 3 (1992/93), pp. 59-90.
[3] Keren Yarhi-Milo, *Knowing the Adversary: Leaders, Intelligence, and Assessments of Intentions in International Relations* (Princeton, N.J.: Princeton University Press, 2014).
[4] Eric Min, *Words of War: Negotiation as a Tool of Conflict* (Ithaca, N.Y.: Cornell University Press, 2025).
[5] Kenneth Waltz, *Theory of International Politics* (New York: McGraw-Hill, 1979); Charles Glaser, *Rational Theory of International Politics* (Princeton, N.J.: Princeton University Press, 2010).
[6] Dominic D. P. Johnson, *Overconfidence and War: The Havoc and Glory of Positive Illusions* (Cambridge, Mass.: Harvard University Press, 2004).
[7] Geoffrey Blainey, *The Causes of War*, 3rd ed. (New York: Free Press, 1988).
[8] Lawrence Freedman, "Political Impatience and Military Caution," *Journal of Strategic Studies* 44, no. 1 (2021), pp. 91-116; Gregory F. Treverton, "Theory and Practice," *Intelligence and National Security* 33, no. 4 (2018), p. 477.
[9] Jonathan Kirshner, *An Unwritten Future: Realism and Uncertainty in International Politics* (Princeton, N.J.: Princeton University Press, 2022), pp. 51-67.
[10] Richard K. Betts, "Is Strategy an Illusion?" *International Security* 25, no. 2 (2000), pp. 5-50. The same caveat applies to assessments of uncertainty about current states of the world, where the fact that a judgment turned out to be wrong does not mean it was necessarily unreasonable: see, for example, Robert Jervis, "Reports, Politics, and Intelligence Failures: The Case of Iraq," *Journal of Strategic Studies*, Vol. 29, no. 1 (2006), pp. 3-52.

chance of success, then we can see whether generals actually win those battles roughly seventy percent of the time.[11] Yet that approach is difficult to implement in national security affairs, where important events are rare enough to make statistical evaluation challenging, practitioners rarely make explicit assessments of uncertainty, and the most important judgments are often classified.[12]

To understand how cognitive biases might influence national security decision-making, scholars frequently analyze "non-elite" samples, such as college students or participants recruited from the general population. For example, the "Good Judgment Project" conducted a multi-year study that recruited thousands of people to make nearly one million forecasts about international politics.[13] These predictions tended to be overconfident, but it is hard to extrapolate the extent to which this finding generalizes to national security officials. Some studies show that elite and non-elite populations exhibit similar psychological tendencies;[14] others argue that national security officials should display fewer cognitive biases than the broader public;[15] others find that national security officials have biases that are *not* prominent among non-elites;[16] others theorize that overconfidence is one bias that elites and non-elites are especially likely to share.[17] It is not possible to resolve these debates without conducting large-scale analyses of how well (or how poorly) national security officials assess uncertainty.

This study aims to answer these questions by analyzing a novel data set containing over 60,000 assessments of uncertainty made by nearly 2,000 military and civilian national security officials from more than forty NATO allies and partners. This is by far the largest publicly available body of probability estimates made by national security practitioners. It is also the only large-scale study of its kind that spans military and civilian elites from many national backgrounds. These data have important limitations: most notably, they were gathered by asking national security officials to answer surveys rather than by analyzing the output of structured analytic processes. The findings are thus primarily useful for evaluating national security officials' intuitive abilities to assess uncertainty; they do not reflect real-world judgments. Yet the next section explains that there are many reasons to expect that intuitive biases shape national security decisions. It is important for national security bureaucracies to identify and mitigate those flaws.

---

[11] Philip E. Tetlock, *Expert Political Judgment: What Is It? How Can We Know?* (Princeton, N.J.: Princeton University Press, 2005).

[12] Jeffrey A. Friedman, *War and Chance: Assessing Uncertainty in International Politics* (New York: Oxford University Press, 2019), pp. 17-50. Furthermore, many national security outcomes do not lend themselves to clear classifications: in many cases, the boundary between "success" and "failure" is subjective, which makes it even harder to judge whether decision-makers were over- or underoptimistic when making high-stakes choices.

[13] Barbara A. Mellers et al., "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," *Perspectives on Psychological Science* 10, no. 3 (2015), pp. 267-281; Philip E. Tetlock and Daniel Gardner, *Superforecasting: The Art and Science of Prediction* (New York: Broadway, 2015).

[14] Joshua D. Kertzer, "Re-Assessing Elite-Public Gaps in Political Behavior," *American Journal of Political Science* 66, no. 3 (2022), pp. 539-553.

[15] William H. Riker, "The Political Psychology of Rational Choice Theory," *Political Psychology* 16, no. 1 (1995), pp. 23-44.

[16] Alex Mintz, Steven B. Redd, and Arnold Vedlitz, "Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?" *Journal of Conflict Resolution*, Vol. 50, no. 5 (2006), pp. 757-776.

[17] Emilie M. Hafner-Burton, D. Alex Hughes, and David G. Victor, "The Cognitive Revolution and the Political Psychology of Elite Decision Making," *Perspectives on Politics* 11, no. 2 (2013), pp. 368-386.

The data reveal that national security officials' intuitions are overwhelmingly overconfident.[18] For example, when study participants estimated that statements had a ninety percent chance of being true, those statements were true just fifty-eight percent of the time. Ninety-six percent of participants would have improved their performance if they had made every one of their judgments with less certainty. In short, if you are a national security professional, the world is probably more uncertain than you think.[19]

This finding extends an emerging body of scholarship showing how foreign policy practitioners share cognitive biases that are widespread among the general public.[20] For example, this study indicates that national security officials were significantly more overconfident than participants in the Good Judgment Project, and that this bias was comparable to the results of identical surveys administered to respondents on Amazon Mechanical Turk.[21] This pattern is remarkable given that national security bureaucracies have strong incentives to cultivate skills for assessing uncertainty accurately.[22] Yet, most national security bureaucracies do not systematically gather data to identify and correct judgmental biases.[23] This article shows that it would be feasible and desirable to implement such procedures.

The study's findings provide several additional insights for policy and scholarship. For example, experimental evidence shows that just two minutes of training significantly reduced national security officials' overconfidence. National security officials' cognitive biases are thus widespread, but they also appear to be tractable if national security organizations are willing to combat them with relatively small amounts of effort. Another experiment embedded within this study demonstrates that national security officials' intuitions for assessing uncertainty are especially prone to false positives. This implies that there may be a shared cognitive foundation for several phenomena that scholars typically treat as distinct, such as mutual optimism in war (in which both sides overestimate their chances of success), threat inflation (which involves attaching excessive certainty to ambiguous claims), and overrating the probability of changes to the status quo. Understanding that national security officials' judgments are prone to false positives also

[18] As noted above, this article uses the term "overconfidence" to describe individuals who assign too much certainty to their judgments. Other scholars sometimes use the term "overconfidence" to describe other attributes, such as individuals who overestimate their capabilities at performing a task. See, for example, Pietro Ortoleva and Erik Snowberg, "Overconfidence in Political Behavior," *American Economic Review*, Vol. 105, no. 2 (2015), pp. 504-535; Dominic D. P. Johnson et al., "Overconfidence in Wargames: Experimental Evidence on Expectations, Aggression, Gender, and Testosterone," *Proceedings of the Royal Society B*, Vol. 273, no. 1600 (2006), pp. 2513-2520.

[19] The phrase "the world is more uncertain than you think" is borrowed from Richard Zeckhauser's "analytic maxims." See Dan Levy, *Maxims for Thinking Analytically: The Wisdom of Legendary Harvard Professor Richard Zeckhauser* (Cambridge, Mass.: Dan Levy, 2021).

[20] Kertzer, "Re-Assessing Elite-Public Gaps in Political Behavior."

[21] Jeffrey A. Friedman, Jennifer S. Lerner, and Richard Zeckhauser, "Behavioral Consequences of Probabilistic Precision," *International Organization*, Vol. 71, no. 4 (2017), pp. 803-826.

[22] Peter J. Katzenstein and Lucia A. Seybert eds., *Protean Power: Exploring the Uncertain and Unexpected in World Politics* (New York: Cambridge University Press, 2018); Jennifer E. Sims, *Decision Advantage: Intelligence in International Politics from the Spanish Armada to Cyberwar* (New York: Oxford University Press, 2022).

[23] The U.S. Intelligence Community, for example, has traditionally resisted proposals to gather systematic data on the accuracy of its judgments, on the grounds that this information could expose analysts to excessive criticism. See Stephen Marrin, "Evaluating the Quality of Intelligence: By What (Mis)Measure?" *Intelligence and National Security* 27, no. 6 (2012), pp. 896-912. In conducting research for this project at institutions of professional military education in several countries, every cohort of participants noted that this kind of feedback – and, indeed, the idea that it was even possible to gather and analyze such information – was novel to them.

carries practical implications—suggesting, for example, that intelligence agencies should avoid "single-outcome forecasting" by ensuring that analysts always consider multiple hypotheses when assessing uncertainty.[24]

Finally, the data reveal that national security officials display similar cognitive biases regardless of whether they are asked to assess uncertainty about the future versus current issues, and regardless of whether they are asked to express their judgments using numbers or words. This suggests that surveys eliciting numeric assessments of uncertainty about factual matters, which can be conducted in minutes, reveal biases that are relevant for understanding how national security officials make forecasts using natural language, which can take months or years to process. These patterns provide additional evidence that national security bureaucracies can leverage insights from decision science to improve cognitive performance at scale.[25]

Studying cognitive biases among national security officials

How well (or how poorly) do national security officials assess uncertainty? The most rigorous studies of this subject offer mixed answers to that question. For example, when David Mandel and Alan Barnes examined 1,514 forecasts made by the Canadian Intelligence Secretariat's Middle East Division, they found that those judgments were systematically underconfident.[26] When Nicholas Miller analyzed 199 judgments from U.S. National Intelligence Estimates regarding nuclear proliferation, he found that those judgments were initially quite overconfident, but that the quality of these assessments improved over time.[27] When Bradley Stastny and Paul Lehner analyzed 99 forecasts made by U.S. intelligence analysts on a range of subjects, they found that those judgments were overconfident in some areas, underconfident in others, and poorly calibrated on the whole.[28] Each of these studies offers important contributions, particularly in showing that scholars can rigorously evaluate assessments of uncertainty in national security contexts. Each nevertheless examines a relatively small volume of data drawn from relatively narrow subsets of practitioners. It is thus unsurprising that these studies reach conflicting conclusions. All of these studies, moreover, focus on civilian intelligence analysts, whose behavior may not generalize to national security professionals writ large.

Other scholars have examined the challenges of assessing uncertainty in international politics by drawing study participants from the general population. The Good Judgment Project, for example, recruited more than 2,000 individuals to make geopolitical forecasts.[29] It identified a group of "superforecasters" who made highly accurate predictions, but found that study participants were, on the whole, moderately overconfident: for example, when Good Judgment

[24] Willis C. Armstrong, William Leonhardt, William J. McCaffrey, and Herbert C. Rothenberg, "The Hazards of Single-Outcome Forecasting," *Studies in Intelligence* 38, no. 3 (1984), pp. 57-70.
[25] Cf. Rose McDermott, "Experimental Intelligence," *Intelligence and National Security* 26, no. 1 (2011), pp. 82-98; Mandeep K. Dhami, Barbara A. Mellers, and Philip E. Tetlock, "Improving Intelligence Analysis with Decision Science," *Perspectives on Psychological Science* 10, no. 6 (2015), pp. 753-757.
[26] David R. Mandel and Alan Barnes, "Accuracy of Forecasts in Strategic Intelligence," *PNAS* 111, no. 30 (2014), pp. 10984-10989; David R. Mandel, "Accuracy of Intelligence Forecasts from the Consumer's Perspective," *Policy Insights from the Behavioral and Brain Sciences* 2, no. 1 (2015), pp. 111-120.
[27] Nicholas L. Miller, "Learning to Predict Proliferation," *International Organization* 76, no. 2 (2022), pp. 487-507.
[28] Bradley J. Stastny and Paul E. Lehner, "Comparative Evaluation of Forecast Accuracy of Analysis Reports and a Prediction Market," *Judgment and Decision Making* 13, no. 2 (2018), pp. 202-211.
[29] Mellers et al., "Identifying and Cultivating Superforecasters"; Tetlock and Gardner, *Superforecasting*.

Project forecasters estimated that an outcome had a ninety-five percent chance of taking place, those outcomes occurred closer to eighty-five percent of the time.[30] Yet, as noted earlier, it is not obvious that this finding applies to national security professionals, who have more incentives than the general population to hone their ability to assess uncertainty, who devote their careers to studying world politics, and who possess unique professional cultures that might encourage excessive caution rather than overconfidence. In sum, no empirical study to date provides generalizable foundations for understanding the extent to which national security officials' assessments of uncertainty are systematically biased in one direction or another.

To tackle that challenge, this study partnered with four advanced military education programs: the Canadian Forces College, the NATO Defense College, the Norwegian Defence Intelligence School, and the U.S. National War College.[31] These institutions comprise large, diverse samples of national security professionals. In Canada, Europe, and the United States, military officers who obtain the rank of colonel are normally required to complete a graduate degree at these kinds of institutions. The NATO Defense College and the U.S. National War College serve an especially diverse range of countries, drawing students from more than forty NATO allies and partners.[32] These institutions' cohorts also contain substantial numbers of civilian national security officials drawn from foreign affairs ministries, intelligence agencies, and other areas of government tasked with responsibilities related to international affairs.[33] These institutions agreed to administer online surveys as part of their core curricula in exchange for providing participants individualized feedback about their cognitive biases. Participation rates exceeded ninety percent for most cohorts. A total of 1,894 national security officials participated in this exercise.[34] These officials made 63,130 assessments of uncertainty.

This study design has several advantages over prior research. For example, the study contains roughly thirty times as many assessments of uncertainty as Mandel and Barnes' analysis of Canadian intelligence officials, which was previously the largest publicly-available data set

---

[30] Barbara Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," *Psychological Science*, Vol. 25, no. 5 (2014), p. 1112.

[31] The study was administered across nineteen sessions from 2015-2023. The Appendix details the composition of each survey cohort. The study was approved by the Dartmouth College Committee for the Protection of Human Subjects, Study #28925.

[32] For example, the 2022 class of NATO Defence College students who participated in the study contained 108 students from thirty-four countries: Algeria (1 student), Armenia (2), Azerbaijan (2), Belgium (1), Canada (1), Denmark (2), Egypt (3), France (6), Georgia (2), Germany (8), Greece (2), Hungary (2), Iraq (4), Italy (11), Jordan (3), Kuwait (5), Mauritania (3), Moldova (1), Mongolia (2), Morocco (3), the Netherlands (2), Norway (4), Pakistan (1), Poland (1), Saudi Arabia (3), Slovakia (1), Slovenia (1), South Korea (1), Spain (8), Taiwan (1), Tunisia (3), Turkey (6), the United Kingdom (5), and the United States (7). The U.S. National War College also serves national security officials from a wide range of nationalities which, in addition to those listed above, include Afghanistan, Bosnia and Herzegovina, Chile, Israel, Mexico, and Sweden. For the purposes of maintaining anonymity, surveys did not ask respondents to declare their nationality, as this information would have been sufficient to identify many individuals. Participating institutions did allow the survey to ask whether participants were U.S. citizens, and those comprised fifty-nine percent of the overall sample. Yet, the Appendix shows that U.S. citizens who participated in this study were overwhelmingly drawn from the U.S. National War College, while comprising less than ten percent of respondents from the other three participating institutions. The Appendix also shows that all of this study's findings hold when analyzing data from each institution individually. Thus, while demands for ensuring respondent anonymity preclude granular demographic analysis, we can be confident that the study's results reflect patterns that hold across national security officials drawn from a wide range of nationalities.

[33] Seventy-three percent of study participants were active-duty military and 27 percent were civilian.

[34] Eighty-four percent of participants in this study were men.

examining national security officials' probabilistic judgments.[35] Whereas most prior studies of this subject involve relatively narrow samples of personnel, often drawn from one office within one country and almost always focusing on intelligence analysts specifically, this study involves a wide variety of civilian and military officials who represent a wide range of nationalities. While it is, of course, impossible to know whether this study's findings apply to states, such as China, that do not send national security officials to institutions associated with NATO, we can at least be confident that the cognitive biases documented in this article generalize broadly—that they are not the product of particular countries or institutional cultures. And, while survey research on national security elites often suffers from low participation rates that raise questions about representativeness,[36] the data described in this article reflect judgments made by nearly every national security official who was assigned to one of the educational programs with which the study partnered.[37]

Each survey asked participants to estimate the chances that thirty to forty statements were true. These questions were regularly updated across survey waves, covering a variety of topics related to international military, economic, and political affairs. In total, the study contained more than 250 unique questions. Every survey was cleared in advance by participating institutions to ensure that its content was relevant to the national security officials with whom they worked.

Most questions asked respondents to assess uncertainty about current issues, such as "In your opinion, what are the chances that NATO's members spend more money on defense than the rest of the world combined?"[38] Assessments of uncertainty on these questions could be evaluated immediately, in order to give national security professionals feedback as soon as the survey concluded.[39] Other questions asked participants to make forecasts that could only be evaluated at later dates, such as "In your opinion, what are the chances that Russia and Ukraine will officially

---

[35] Mandel and Barnes, "Accuracy of Forecasts in Strategic Intelligence"; Mandel, "Accuracy of Forecasts from the Consumer's Perspective."

[36] Simone Dietrich, Heidi Hardt, and Haley J. Swedlund, "How to Make Elite Experiments Work in International Relations," *European Journal of International Relations* 27, no. 2 (2021), pp. 596-621.

[37] While national security officials are generally required to study at professional military education institutions as a condition for promotion, participation in this study was voluntary. Incomplete surveys (which were rare) were dropped from the sample in order to diminish concerns about "biased missingness" in the data.

[38] Other questions included asking respondents to estimate the chances that NATO had currently deployed more than 15,000 troops to Afghanistan, whether more than 30 countries currently participated in China's Belt and Road Initiative, whether Saudi Arabia currently exports more oil than all other countries combined, and whether there are currently more refugees from Syria or Venezuela.

Since questions about current states of the world all had "right answers," fully-informed participants could have answered all of them with estimates of zero or one hundred percent. Yet, most participants did not know the answers to most questions posed, and thus needed to provide their personal degrees of belief that the statements were true. This exercise in estimating subjective probability is equivalent to the challenge national security officials face when confronting imperfect information about current states of the world. For example, when national security officials considered the chances that Iraq was pursuing nuclear weapons in 2002 or the chances that Osama bin Laden was living in Abbottabad in 2011, their conclusions reflected personal degrees of belief in statements that were, in reality, either true or false.

[39] Posing factual questions in the survey raised the possibility that some participants might look up the right answers, which the survey instructed them not to do. If anything, the prospect of noncompliance would make it harder for the survey to document cognitive biases, as noncompliance would have increased the accuracy of participants' responses.

declare a ceasefire by the end of 2022?"[40] As shown below, national security professionals demonstrated similar cognitive biases across these question formats.

Most assessments of uncertainty in the study were elicited as numeric percentages, which made it possible to give clear feedback to the participants regarding their judgmental biases. However, quantitative assessments of uncertainty might seem inapt, given that national security officials often express uncertainty using qualitative language. To address this issue, a random subset of responses was elicited using qualitative terms, such as "likely" and "almost certain," that are recommended for use in the U.S. Intelligence Community.[41] This variation also had no meaningful impact on results.

This study's primary drawback is that national security officials naturally invest less effort into completing surveys than they would devote to analyzing real decisions. This limitation is essentially unavoidable for experimental research on high-stakes decision-making.[42] Results should thus be interpreted as measuring participants' *intuitions* for assessing uncertainty, recognizing that these intuitions are just one input to national security analysis and decision-making. As Daniel Kahneman might phrase it, these data reflect national security officials "thinking fast"—the data presented below capture national security officials' "cognitive first steps" when assessing uncertainty.[43]

These intuitions matter for two main reasons. First, there is substantial evidence that individuals' initial, intuitive impressions of a problem anchor their subsequent judgments.[44] Even if deliberative analysis can mitigate the impact of intuitive cognitive errors, the first steps that national security professionals take when assessing uncertainty shape their subsequent performance. This argument is consistent with findings from Joshua Kertzer and colleagues showing that individual-level cognitive biases persist in group settings.[45] Even if group deliberation often improves analytic rigor, it does not necessarily eliminate flaws in human judgment. In some cases, group deliberation can *enhance* cognitive biases—for example, by suppressing heterodox viewpoints[46] or through herding behavior that encourages individuals to adopt more extreme views.[47]

---

[40] Other examples included asking participants to estimate the chances that more than 10 U.S. soldiers would be killed fighting ISIS within the next six months, whether Iraqi Security Forces would reclaim control of Ramadi or Mosul within six months, whether Liz Truss would be elected as Britain's next Prime Minister, whether NATO would ratify membership for Finland and Sweden by the end of 2022, and whether that year would be the hottest year on record.

[41] Mandeep K. Dhami and David R. Mandel, "Words or Numbers? Communicating Probability in Intelligence," *American Psychologist* 76, no. 3 (2021), pp. 549-560.

[42] Alex Mintz, Yi Yang, and Rose McDermott, "Experimental Approaches to International Relations," *International Organization* 76, no. 2 (2011), pp. 493-501.

[43] Daniel Kahneman, *Thinking, Fast and Slow* (New York: FSG, 2011).

[44] Robert Jervis 1976, *Perception and Misperception in International Politics* (Princeton, N.J.: Princeton University Press, 1976), pp. 143-202; Nicholas Epley and Thomas Gilovich, "The Anchoring-and-Adjustment Heuristic," *Psychological Science* 17, no. 4 (2006), pp. 311-318.

[45] Joshua D. Kertzer, Marcus Holmes, Brad L. LeVeck, and Carly Wayne, "Hawkish Biases and Group Decision Making," *International Organization* 76, no. 2 (2022), pp. 513-548.

[46] Irving L. Janis, *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascos* (Boston, Mass.: Houghton Mifflin, 1972).

[47] Carly Wayne, Mitsuru Mukaigawara, Joshua D. Kertzer, and Marcus Holmes, "Diplomacy by Committee: Assessing Resolve and Costly Signals in Group Settings," *American Journal of Political Science*, in press.

Additionally, national security officials often make high-stakes choices under conditions of stress and time scarcity that preclude the use of structured analytic processes. These constraints, which are essentially unavoidable in tactical decision-making, force individuals to rely on their intuitions in a manner that amplifies the effects of cognitive biases.[48] Even at strategic levels, national security officials frequently form beliefs based on intuitions rather than on conducting extensive deliberations or reading rigorous intelligence reports.[49] For example, high-ranking members in the George W. Bush administration devoted little systematic effort to assessing the long-term risks of invading Iraq.[50] Instead, they based their decision to go to war on intuitive assumptions that the U.S. military could easily stabilize Iraq after toppling Saddam Hussein's regime.[51] This is just one salient example of why it is important to understand the accuracy of national security officials' intuitions for assessing uncertainty, to identify the biases that those intuitions contain, and to determine whether those problems can be mitigated.[52]

National security officials' intuitions are overwhelmingly overconfident

Figure 1 presents a "calibration curve" describing the 50,408 numeric assessments of uncertainty that this study collected.[53] The figure's horizontal axis captures the chances that national security officials assigned to statements being true. The vertical axis indicates the proportion of the time that those statements were actually true. If national security officials' intuitions for assessing uncertainty were well-calibrated, then the data would fit a 45-degree line, such that when study participants said a statement had a thirty percent of being true, then those statements would actually be true thirty percent of the time.

Instead, Figure 1 reveals that the national security officials who participated in this study were overwhelmingly overconfident. For instance, when officials thought there was a ninety percent chance that a statement was true, those statements were true just fifty-seven percent of the time. This degree of overconfidence is at least as large as what other studies have previously documented in non-elite samples. For example, the national security officials who contributed to this study were significantly more overconfident than forecasters who participated in the Good Judgment

[48] Gary Klein, *Seeing What Others Don't: The Remarkable Ways We Gain Insights* (New York: PublicAffairs, 2013).
[49] Yarhi-Milo, *Knowing the Adversary*. It is thus not obvious whether we should expect national security officials' intuitions to be more impactful at strategic versus tactical levels, overall: strategic analyses often involve less time pressure, but national security officials who work at this level may also tend to conduct less disciplined debates. It would be highly unusual, for example, for national security principals to work through the kinds of "structured analytic techniques" that are widely employed by rank-and-file intelligence analysts.
[50] Aaron Rapport, *Waging War, Planning Peace: U.S. Noncombat Operations and Major Wars* (Ithaca, N.Y.: Cornell University Press, 2015), pp. 82-123.
[51] Melvyn P. Leffler, *Confronting Saddam Hussein: George W. Bush and the Invasion of Iraq* (New York: Oxford University Press, 2023), pp. 149-202.
[52] Emilie M. Hafner-Burton, Stephan Haggard, David A. Lake, and David G. Victor, "The Behavioral Revolution and International Relations," *International Organization* 71, S (2017), pp. S1-S31.
[53] Qualitative assessments of uncertainty gathered in this study are analyzed below. All curves plotted on graphs in this article reflect local polynomials with 95 percent intervals.

Project, and they were roughly as overconfident as a group of 775 respondents recruited to take the same survey that was administered to one of the study's National War College cohorts.[54]

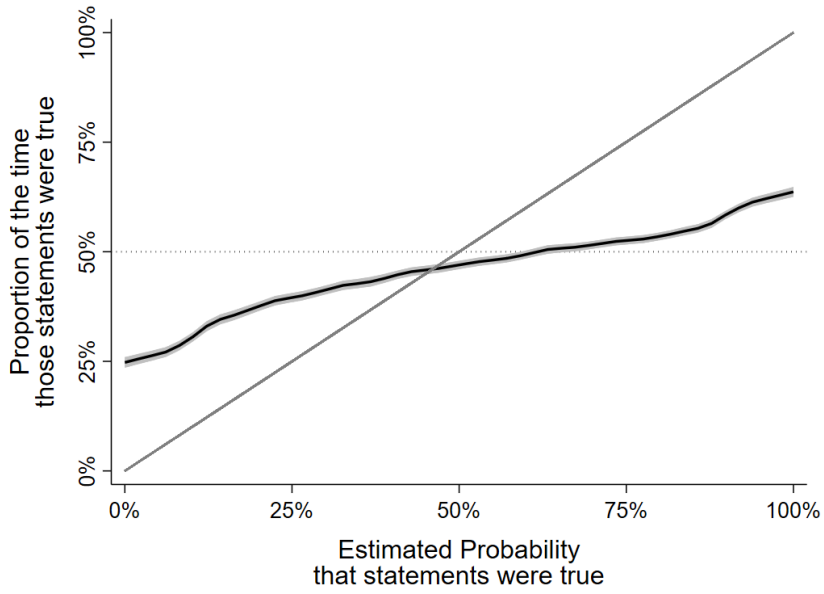

Figure 1. *Judgmental calibration for 50,408 assessments of uncertainty made by national security professionals*.

This pattern of overconfidence was remarkably consistent across the data. All nineteen cohorts of national security officials who participated in the study gave overconfident estimates. This bias appeared for both civilian and military professionals, for both men and women, and for both U.S. and non-U.S. citizens.[55] This bias also appeared across a wide range of subject matter: limiting the

[54] For example, Figure 1 shows that the statements to which national security officials assigned an 80 percent probability were true just 55 percent of the time, and that the statements to which national security officials assigned a 95 percent probability were true just 62 percent of the time. These outcome frequencies were 75 percent and 85 percent, respectively, for the Good Judgment Project and 55 percent and 55 percent, respectively, for assessors on Amazon Mechanical Turk, who took the same survey administered at the National War College in 2015. For Good Judgment Project calibration data, see Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," p. 1112. Data from Amazon Mechanical Turk are contained in replication materials for Friedman, Lerner, and Zeckhauser, "Behavioral Consequences of Probabilistic Precision."

[55] The Appendix shows that there were no statistically-significant differences in performance when dividing the sample between men/women or by military/civilian status. U.S. citizens displayed marginally less overconfidence than U.S. citizens, but this difference was substantively small (roughly one-fifth of a standard deviation) and it is likely associated with the fact that U.S. citizens may have found it easier to engage with an English-language survey. The study's findings are consistent with those from the Good Judgment Project, which also found no significant differences in performance between men and women: see Mark Himmelstein, Pavel Atanasov, and David V. Budescu, "Forecasting Forecaster Accuracy: Contributions of Past Performance and Individual Differences," *Judgment and Decision Making*, Vol. 16, no. 2 (2021), pp. 339, 349.

analysis to virtually any subset of survey questions produces similar results. (See the Appendix for details.)

The data also show that national security officials' judgments were biased towards false positives. Figure 1 documents this pattern by showing that national security officials' assessments were particularly overconfident when they estimated probabilities above fifty percent. We can quantify this overconfidence by measuring the difference between the probabilities that national security officials assigned to their judgments and the actual proportion of those claims that were true. Thus, when the statements to which participants assigned a ninety percent probability turned out to be true just fifty-seven percent of the time, that represents a bias of thirty-three percentage points. By contrast, if we look at statements to which participants assigned a ten percent probability—a degree of certainty that is logically equivalent to judgments of ninety percent—those statements turned out to be true thirty-two percent of the time, for a gap of twenty-two percentage points. In other words, national security officials appear to have a particularly tendency to believe that false statements are true. Later sections of this article will present further evidence to document that bias and explain why it has important implications for the theory and practice of national security decision-making.

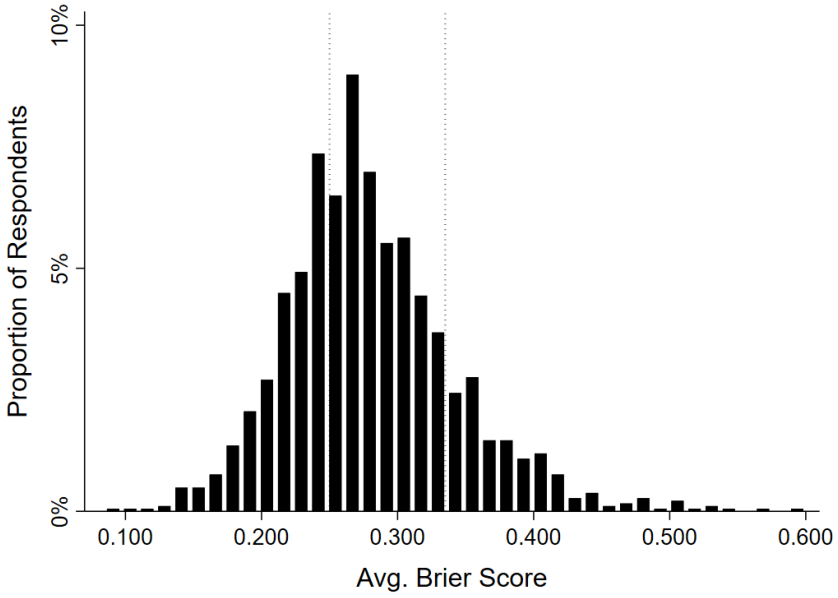

Figure 2. *Brier Scores for 1,470 national security professionals*.

Figure 2 quantifies the average accuracy of each study participant's judgments using "Brier Scores." Brier Scores capture the squared difference between the probability estimates an individual made and the estimates they could have made if they knew each question's "right answer" with certainty.[56] Since Brier Scores measure squared error, lower numbers indicate more

[56] Thus, if an individual assigns a probability of 0.75 to a statement that is true, then the Brier Score for that judgment is $(1.0\text{-}0.75)^2=0.0625$. The Brier Score is the most common metric that scholars have previously used for evaluating assessments of uncertainty in international politics: see, for example, Tetlock, *Expert Political Judgment* and Mellers et al., "Identifying and Cultivating Superforecasters."

accurate judgments. The vertical lines in Figure 2 reflect two benchmarks for gauging performance. A Brier Score of 0.250 is the score that participants would have received if they claimed complete ignorance, and thus recorded probability estimates of fifty percent, for every question the survey posed. A Brier Score of 0.335 is the score that participants would have received, on average, if they had responded to the survey by making probability estimates at random.

The average participant's Brier Score in this study was 0.280,[57] with sixty-eight percent of participants receiving Brier Scores that were worse than 0.250. In other words, most national security officials in this study would have performed better if they simply said they did not know the answer to every question that the survey gave them. Sixteen percent of participants would have received better scores, in expectation, if they had guessed probabilities at random.

These findings do not indicate that national security officials lack knowledge or that they cannot think probabilistically. Figure 1 clearly demonstrates that study participants had reliable intuitions for judging which statements were more likely to be true than others.[58] Yet national security officials' overconfidence was so extreme that it essentially canceled out the knowledge that these individuals possessed. Ninety-six percent of national security officials who participated in the study would have received better Brier Scores if they had attached less certainty to every one of their judgments.[59]

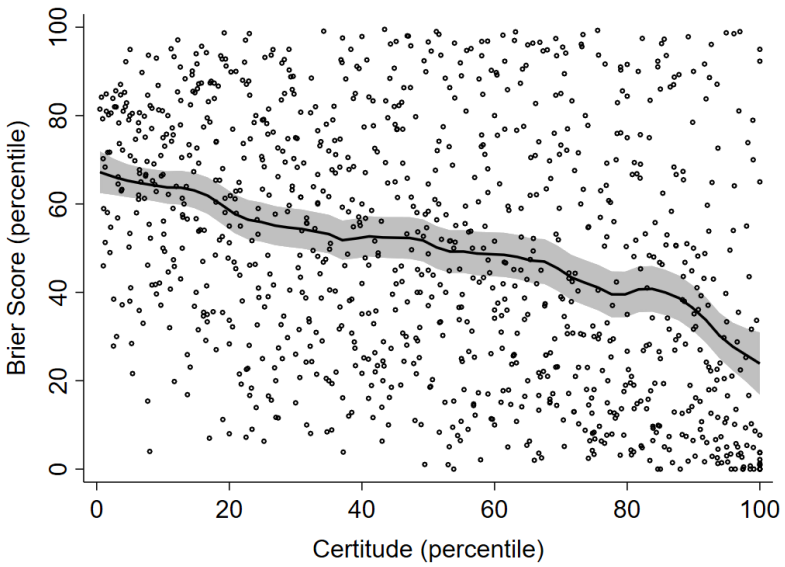

Figure 3. *National security officials who assigned more certainty to their judgments also tended to be less accurate.*

[57] Standard deviation 0.063.
[58] This skill is known as judgmental "discrimination." For more evidence that foreign policy analysts can display excellent judgmental discrimination despite poor judgmental calibration, see Jeffrey A. Friedman et al., "The Value of Precision in Geopolitical Forecasting," *International Studies Quarterly* 62, no. 2 (2018), pp. 410-422.
[59] See Appendix for documentation.

Figure 3 indicates that participants who attached less certainty to their judgments tended to be more accurate, overall. The horizontal axis of this graph reflects each national security official's "certitude": the average distance between their assessments of uncertainty and fifty percent.[60] The graph's vertical axis captures each national security official's Brier Score. Figure 3 "normalizes" these attributes into percentile rankings within each survey cohort in order to minimize confounds that might result from different groups receiving different questions at different times. The graph reveals a consistent, negative relationship between certitude and judgmental accuracy.[61] In other words: the more certainty national security officials possessed in this study, the less accurate their judgments tended to be.

Mitigating overconfidence through brief feedback

Showing that national security officials are overwhelmingly overconfident does not imply that their biases are impossible to correct. It is plausible that their overconfidence stems, at least in part, from the fact that most national security officials do not receive explicit feedback about their abilities to assess uncertainty. In the absence of such feedback, it is easy to develop "illusions of skill."[62] Philip Tetlock, for example, has documented a tendency for experts to give themselves full credit for making judgments that seem wise after the fact while "explaining away" their failures in a manner that prevents effective learning.[63] How hard is it to burst these illusions and thereby improve performance?

The standard tool that decision scientists deploy for this purpose is called "calibration training." This method involves asking participants to assess uncertainty, providing feedback on the accuracy of their judgments, and then administering follow-up surveys to measure improvement over time.[64] That approach was infeasible in the context of this research, where national security officials were only available to take a single survey.

This study thus took a different approach to combating cognitive biases by providing a random subset of individuals with information at the start of each survey describing the biases that national security officials had previously demonstrated in prior surveys.[65] This feedback explained that prior participants' judgments were systematically overconfident, documented that claim by presenting a calibration curve like the graph in Figure 1, and explained that almost all participants would have achieved better scores if they had assigned less certainty to every one of their judgments. (See the Appendix for details.) This extra information was not demanding: on average, participants spend two minutes reading it before moving on to the remainder of the survey.

---

[60] Thus, if a national security official assigned complete certainty to every judgment, their average certitude would be 0.50; if another individual assigned a probability of either 25 percent or 75 percent to every statement in the study, their average certitude would be 0.25.

[61] This relationship is highly statistically significant ($p<0.001$).

[62] Kahneman, *Thinking, Fast and Slow*, pp. 216-217.

[63] Philip E. Tetlock, "Theory-Driven Reasoning about Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions?" *American Journal of Political Science* 43, no. 2 (1999), pp. 335-366.

[64] Sarah Lichtenstein, Baruch Fischhoff, and Paul Slovic, "Calibration of Probabilities" in Daniel Kahneman, Paul Slovic, and Amos Tversky eds., *Judgment Under Uncertainty* (New York: Cambridge University Press, 1982), pp. 294-305.

[65] 689 participants received this information versus 643 who did not.

Armed with this feedback, national security officials made much better assessments of uncertainty. They posted average Brier Scores of 0.274, while participants in the control group posted average Brier Scores of 0.291. This improvement was highly significant in both statistical and substantive terms.[66] National security officials were similarly receptive to feedback regardless of whether they were men or women, military or civilian personnel, and U.S. or non-U.S. citizens. As expected, improved performance was associated with the fact that participants who received feedback about prior cohorts' overconfidence attached less certainty to their judgments.[67] Almost all the improved performance in the treatment group (ninety-one percent) is attributable to the fact that they became more cautious when assessing uncertainty.[68]

This finding is consistent with prior research showing that decision-makers can be trained to combat cognitive biases. For example, Megan Kelly and David Mandel found that instructing intelligence analysts to watch a course comprising six instructional videos significantly improved judgmental accuracy.[69] The Good Judgment Project found that a group of forecasters who were randomly assigned to take a one-hour online training program in reducing cognitive biases performed significantly better than their counterparts.[70] This study complements that literature by showing that interventions need not be extensive or sophisticated to have meaningful impact. If national security officials can systematically improve their judgments by receiving just two minutes of training, then national security bureaucracies may be able to combat overconfidence by institutionalizing similarly simple procedures at large scales.

Bias towards false positives

Study participants' assessments of uncertainty were biased towards false positives. In other words, national security officials appear to find it easier to generate ideas about why a hypothesis might be true than why it might be false. Yet, without experimentally manipulating questions, it is difficult to know whether this pattern represents a consistent cognitive bias, as opposed to spurious features of survey design. Survey questions may have unintentionally been phrased in a manner that skewed participants' judgments.

To address this ambiguity, a subset of surveys randomly selected questions from two mutually exclusive and logically identical alternatives. For example, half of participants might receive a question asking "What are the chances that Boko Haram has killed more civilians than ISIS since 2010?" while the other half would receive "What are the chances that ISIS has killed more civilians than Boko Haram since 2010?" Since these hypotheses are the inverse of one another, the average

[66] This different in average performance was equivalent to one-quarter of a standard deviation in the control group's Brier Scores (sd=0.066) and it was statistically significant at the p<0.001 level.
[67] The average certitude for participants who received feedback was 0.23, versus 0.28 in the control group (p<0.001), a reduction of roughly two-thirds of a standard deviation.
[68] This reduction in certitude accounts for ninety-one percent of the treatment group's improved performance (95% CI: 0.65-1.48) according to the method of mediation analysis proposed by Kosuke Imai, Luke Keele, Dustin Tingley and Teppei Yamamoto, "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies," *American Political Science Review* 105, no. 4 (2011), pp. 765-789.
[69] See, for example, Megan O. Kelly and David R. Mandel, "The Effect of Calibration Training on the Calibration of Intelligence Analysts' Judgments," *Applied Cognitive Psychology* 38, no. 4236 (2024), pp. 1-13.
[70] Welton Chang, Eva Chen, Barbara A. Mellers, and Philip E. Tetlock, "Developing Expert Political Judgment," *Judgment and Decision Making* 11, no. 5 (2016), pp. 509-526.

probability that rational individuals assign to them should sum to one hundred percent.[71] If participants' judgments were skewed towards false positives, then these average estimates would sum to more than one hundred percent.

Figure 4 depicts the average participant's response to each of the two mutually exlusive question variants. Across 280 questions that appeared in this experimental module, the average probabilities participants assigned to each question variant summed to 110 percent. That bias is highly statistically significant[72] and is widespread in the data. The average response to each survey question's two variants summed to more than one hundred percent for 244 of the 280 questions in the experiment. This shows that national security officials' assessments of uncertainty were systematically biased towards false positives, and that this bias generalizes across a wide range of questions rather than being driven by performance on an idiosyncratic subset of issues in the study.

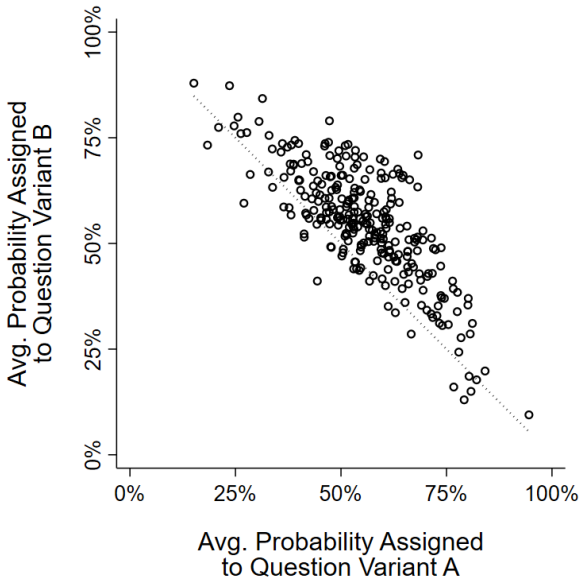

Figure 4. *The average probability that national security officials assigned to mutually exclusive question variants consistently summed to more than 100 percent.*

This bias towards false positives may be related to the "availability heuristic": the tendency for people to exaggerate the chances of outcomes that come more readily to mind.[73] In this view, imagining how a hypothesis might be true may tend to be a more concrete (and thus easier) task than imagining how a hypothesis might be false. If that is the case, then the availability heuristic suggests that most people will have a bias towards confirming, rather than refuting, hypotheses presented to them. The overrepresentation of false positives shown in this survey may also be

[71] Or perhaps slightly lower, given that the two groups could, in principle, have been equally likely.
[72] $p<0.001$.
[73] Amos Tversky and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, Vol. 5, no. 2 (1973), pp. 207-232.

related to "acquiescence bias": the tendency for people to agree with propositions they are asked to consider.[74]

Both of these interpretations have similar relevance for decision-making. In particular, they support the idea that national security officials should avoid the practice of "single-outcome forecasting" in which analysts focus on assessing the chances of a particular hypothesis being true rather than evaluating how uncertainty is distributed among multiple possibilities.[75] A good example of this contrast is how U.S. intelligence analysts studying Iraq's alleged nuclear program in 2002 concluded that Saddam Hussein was importing aluminum tubes in order to build centrifuges for enriching uranium. This argument was plausible, but there was also evidence to indicate that Iraq was using the aluminum tubes to build conventional rockets (an alternate hypothesis that turned out to be correct).[76] Orienting the intelligence process around assessing the chances that Iraq was building nuclear weapons may have exacerbated analysts' bias towards false positives. Simultaneously assessing the chances of multiple hypotheses—in this case the chances that Iraq was using the aluminum tubes for centrifuges, or for rockets, or for some other purpose—can counteract that tendency.[77] Documenting a consistent bias towards false positives also supports prior research arguing that national security officials may benefit from employing a "falsificationist" mindset that explicitly seeks out information to disconfirm statements they think are likely to be true.[78]

Facts versus predictions, words versus numbers

Most assessments of uncertainty in the study pertained to factual matters.[79] Those estimates could be used to provide national security professionals with immediate feedback about their performance.

The task of assessing uncertainty about current states of the world is crucial to many elements of national security decision-making. For example, debates about whether Iraq was pursuing nuclear weapons in 2002, or whether the United States had correctly identified Osama bin Laden's location in 2010, or the state of the U.S.-Soviet nuclear balance during the Cold War, or the extent to which Chinese leaders currently possess revisionist intentions all require assessing uncertainty about factual matters. Yet, national security professionals must also assess uncertainty about future

[74] Jon A. Krosnick, "Survey Research," *Annual Review of Psychology*, Vol. 50 (1999), p 552. While prior research on acquiescence bias has primarily focused on implications for survey research, there are clear analogies to national security analysis. Any time analysts or decision-makers are asked to evaluate the chances that a statement is true—say, the chances that Osama bin Laden is hiding in Abbottabad or the chances that a military operation will succeed—these propositions could potentially stimulate a tendency towards agreement (and, thus, a bias towards false positives).

[75] Armstrong et al., "The Hazards of Single-Outcome Forecasting."

[76] Robert Jervis, *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War* (Ithaca, N.Y.: Cornell University Press, 2010), pp. 127-128, 142-145.

[77] For more discussion on the benefits of assessing how uncertainty is distributed among multiple possibilities—as opposed to making "point estimates" of the chances that a single statement is true—see Jeffrey A. Friedman and Richard Zeckhauser, "Assessing Uncertainty in Intelligence," *Intelligence and National Security*, Vol. 27, no. 6 (2014), pp. 829-834.

[78] Richards Heuer, *The Psychology of Intelligence Analysis* (Washington, D.C. Center for the Study of Intelligence, 1999).

[79] N=61,662.

states of the world: estimating the chances that military operations will succeed, or predicting how another country might respond to diplomatic provocations.[80] To what extent do the study's findings about how national security officials assess factual matters reflect their capabilities for making predictions? To address this question, the study also collected a series of forecasts that were scored at later dates.[81] Figure 5 presents calibration curves for both question types, respectively.

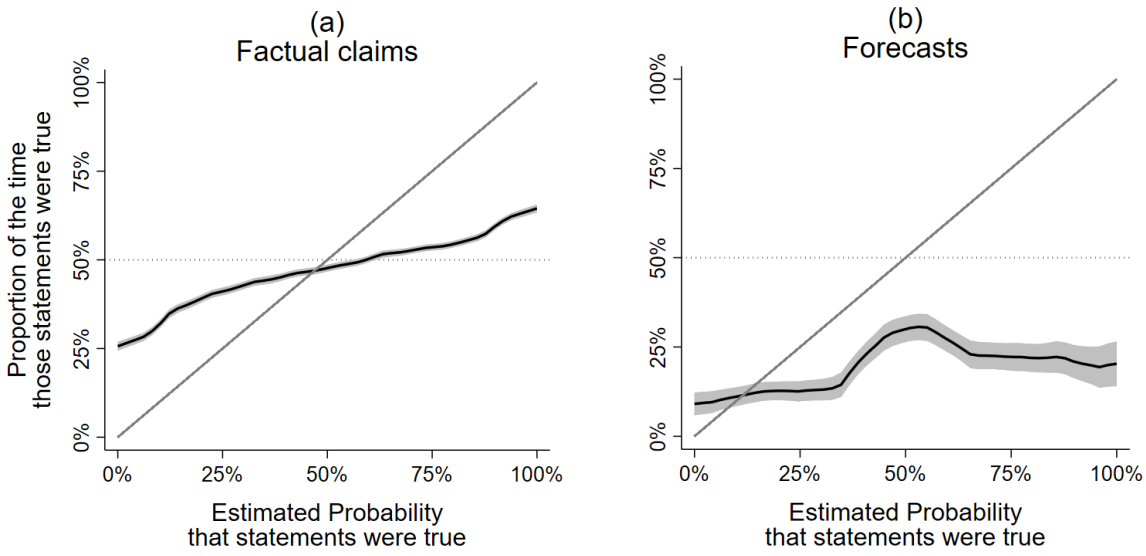

Figure 5. *Judgmental calibration when assessing current versus future states of the world.*

These graphs demonstrate that national security officials' overconfidence, along with their proclivity for false positives, were even more pronounced when making forecasts. Since there is no way to ensure that the surveys' forecasting questions had an equal degree of difficulty to questions regarding current states of the world, Figure 5 cannot sustain causal claims about the degree to which national security professionals' performance differs across these question types. Figure 5 nevertheless shows that this study's findings are not driven by the choice to focus primarily on assessing uncertainty about current states of the world. If anything, it appears that national security bureaucracies have greater reasons to worry about overconfidence and a bias towards false positives when assessing uncertainty about future events.

This study also asked national security officials to assess uncertainty by estimating numeric percentages.[82] As noted earlier, this format facilitates providing clear feedback about judgmental

[80] Intelligence scholars often draw a related distinction between "puzzles," where the right answer would be knowable if analysts possessed the right information, versus "mysteries," where no amount of information could allow reasonable analysts to render judgments with certainty. Gregory F. Treverton, *National Intelligence and Science: Beyond the Great Divide in Analysis and Policy* (New York: Oxford University Press, 2015), pp. 32-35.
[81] *N*=2,546.
[82] *N*=50,408.

biases, but it differs from the way that national security officials often assess uncertainty: verbally and qualitatively. It is possible that asking national security officials to assess uncertainty in unfamiliar ways garbled their thoughts and thereby produced judgmental biases that would not appear in normal settings.[83] In order to test whether this distinction matters, four survey waves assigned a random subset of participants to assess uncertainty using the "words of estimative probability" shown in Figure 6, based on the U.S. National Intelligence Council's then-current guidance for expressing uncertainty.[84]

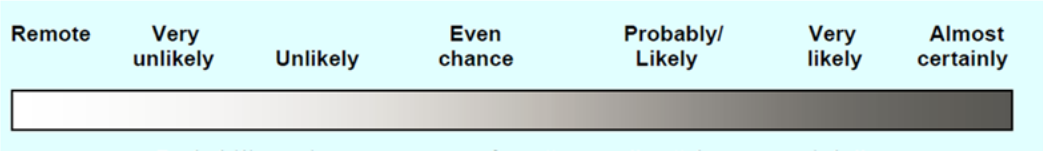

Figure 6. *"Words of estimative probability" lexicon recommended by U.S. National Intelligence Council.*

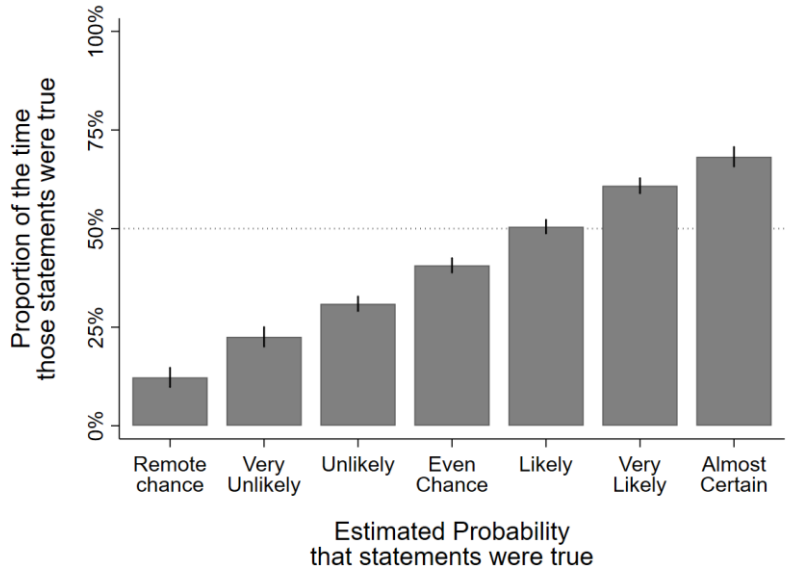

Figure 7. *Judgmental calibration for 13,480 verbal assessments of uncertainty provided by national security officials*.

[83] On the idea that numeric probabilities can elicit biases that do not appear in verbal communication, see Thomas Wallsten, "Costs and Benefits of Vague Information" in Robin M. Hogarth ed., *Insights in Decision Making* (Chicago, Ill.: University of Chicago Press), pp. 28-43; Alf C. Zimmer, "A Model for the Interpretation of Verbal Predictions," *International Journal of Man-Machine Studies*, Vol. 20, no. 1 (1984), pp. 121-134.
[84] *N*=13,480. On the origins and intellectual justifications for this practice, see Sherman Kent, "Words of Estimative Probability," *Studies in Intelligence* 8, no. 4 (1964), pp. 49-65.

Since these words do not carry precise definitions, the accuracy of these data is open to some interpretation. However, national security officials who used these terms were also clearly overconfident. For example, when participants said that a statement was "almost certain" to be true, those statements turned out to be false thirty-two percent of the time (and true the other sixty-eight percent).

Figure 7 also replicates the finding that national security officials' assessments of uncertainty are prone to false positives. For example, when participants said that a statement had a "remote chance" of being true, those statements were true twelve percent of the time—a rate of surprise that was almost three times lower than what national security officials encountered when they assigned the term "almost certain," despite the fact that these judgments reflect equivalent degrees of certitude according to the lexicon used in the experiment.[85]

The article's main findings are thus not particularly sensitive to whether national security officials assessed uncertainty about future versus current states of the world, nor do they depend on eliciting judgments using numbers rather than words. This is welcome news for national security organizations that wish to employ tools from the decision sciences to analyze and improve decision-making. Forecasts that national security officials make using natural language can take years to evaluate. By contrast, surveys eliciting numeric assessments of uncertainty about factual matters can be conducted in a matter of minutes. The fact that such surveys appear to offer generalizable insights about national security officials' cognitive biases provides further evidence that organizations can tackle those problems at scale.

Implications for scholarship and policy

Systematic overconfidence among national security officials has troubling implications for policymaking. National security officials continually make choices that place lives and resources at risk. It is important to minimize those risks wherever possible; indeed, some courses of action are only worth taking if they are almost certain to succeed. Yet, even when national security officials who participated in this study were completely certain that their judgments were right—that is, when they said a statement had either a zero percent or a one hundred percent chance of being true—they were wrong more than one-quarter of the time.[86] If analysts who are completely certain about their conclusions are wrong so often, then leaders must be cautious about trusting anyone who claims that a course of action is truly safe.

These findings also offer implications for international relations theory, particularly with respect to scholarship in the realist tradition that assumes national security officials make rational assessments of uncertainty simply because they have strategic incentives to do so.[87] If national security officials devoted as much effort to bolstering their cognitive capacities as realists assume, then we would not expect their judgments to be so overconfident; we would not expect this bias to be comparable to judgments made by "non-elite" respondents who have no special reasons to

[85] The difference between the proportion of the time that participants who used these terms were surprised by true outcomes (i.e. assigning an "almost certain" judgment to a statement that proved false or a "remote chance" judgment to a statement that proved true) is statistically significant at the $p<0.001$ level.
[86] N=6,835.
[87] See, for example, John J. Mearsheimer and Sebastian Rosato, *How States Think: The Rationality of Foreign Policy* (New Haven, Conn.: Yale University Press, 2023), p. 13.

cultivate talent for assessing uncertainty in world politics; nor would we expect that just two minutes of training would markedly improve national security officials' performance. Each of these findings throws doubt on the assumption that national security officials can reliably assess uncertainty in rational ways.

Relaxing that assumption matters for international relations theory. It suggests that decision-makers are likely to underestimate the risks surrounding national security policies, a bias that is likely to foment international instability.[88] Demonstrating that national security officials' judgments are skewed towards false positives also suggests a cognitive foundation for several phenomena that scholars tend to treat as separate. For example, a proclivity for false positives may be part of the reason why foreign policy analysts tend to overrate the probability of changes to the status quo,[89] and may contribute to threat inflation.[90] The adage that generals "always prepare to fight the last war" is consistent with the idea that, once national security officials have identified a challenge, they will overestimate the chances of encountering that challenge again in the future.[91] Mutual optimism in war—a phenomenon that is widely viewed as a destabilizing force in international politics[92]—also requires at least one state to make a false-positive assessment in overpredicting its chances of obtaining a favorable outcome.

Future research could expand the above analysis in at least two ways. First, study participants could take multiple rounds of surveys.[93] This would provide clearer evidence of the extent to which training produces durable improvements in performance. This procedure would also facilitate experimenting with different training methods in order to determine the most effective approaches to improving assessments of uncertainty.[94] The Good Judgment Project, for example, identified specific training procedures that durably improved the quality of geopolitical forecasts.[95] National security organizations would generally benefit from incorporating such training into professional development programs. It is equally important to understand which interventions have *short*-term impacts that primarily involve "priming" people to think in certain ways without permanently enhancing their skill sets. Bureaucracies would primarily benefit from incorporating these interventions into structured analytic techniques and other standard operating procedures that national security officials employ on a regular basis.

It would also be valuable to gather large-scale data sets that evaluate judgments made by groups, as assessments of uncertainty in national security often reflect corporate judgment rather than

[88] See, for example, Johnson, *Overconfidence and War*.
[89] Tetlock, *Expert Political Judgment*.
[90] Trevor A. Thrall and Jane Kellett Cramer, *American Foreign Policy and the Politics of Fear: Threat Inflation Since 9/11* (New York: Routledge, 2009).
[91] Jack S. Levy, "Learning and Foreign Policy," *International Organization* 48, no. 2 (1994), pp. 279-312.
[92] Blainey, *Causes of War*.
[93] See, for example, Kelly and Mandel, "Effect of Calibration Training."
[94] It could also be worth understanding the conditions under which national security officials are more or less receptive to this feedback. For example, the national security officials in this study may have been unusually receptive to feedback given that they were reached in an educational setting. If this is the case, it suggests that national security organizations should prioritize incorporating material on judgmental calibration into curricula at educational institutions. But if national security officials are just as receptive to feedback in other settings, then these techniques can be applied with a wider reach.
[95] Chang et al. 2016, "Developing Expert Political Judgment."

individual viewpoints.[96] Some studies show that group collaboration can attenuate individual-level cognitive biases, particularly for groups whose members hold diverse viewpoints that expose people to new information they had not previously considered.[97] In other contexts, groupwork has been shown to replicate individual-level biases[98] or even to exacerbate judgmental errors.[99] The latter problem is particularly likely to occur in cases where group members share similar views: here, collaboration runs the risk of amplifying biases in a phenomenon known as "group extremity shift."[100]

In other words, it is not obvious whether we should expect institutional procedures to amplify, mitigate, or maintain national security officials' intuitive overconfidence. Moreover, different kinds of institutional procedures likely interact with cognitive biases in different ways across different contexts.[101] The complex and contingent nature of these relationships suggests that bureaucratic practices likely play an important (and arguably understudied[102]) role in shaping the rationality of national security policy. If national security officials' intuitions for assessing uncertainty are as flawed as this study indicates, then rational decision-making must be mediated by institutional design where possible.

Finally, this study supports some practical advice for national security practitioners. First: remember that the world is more uncertain than you think. Recognize that your intuitions are likely to be overconfident, especially if you have not previously received systematic, quantitative feedback on your assessments of uncertainty. If you think that an outcome is likely to be true, consider those chances to be closer to sixty percent than ninety percent. If you think that an outcome is *un*likely to be true, consider those chances to be closer to forty percent than ten percent.

---

[96] Wargaming, in particular, could provide a valuable platform for determining how group-level judgments may differ from those provided by individuals. Wargaming can also address the question of whether cognitive biases that appear on short surveys generalize to more effortful contexts. The key limitations with using wargames for this purpose is that it may be difficult to randomize key inputs (e.g., individual- versus group-level participation and efforts levels) while holding all other aspects of the wargames equal, and designers would need to embed a very large volume of assessments without wargames in order to generate the volume of data necessary for evaluating judgmental accuracy. On the strengths and limitations of wargames for research on national security decision-making, see Erik Lin-Greenberg, Reid B. C. Pauly, and Jacquelyn G. Schneider, "Wargaming for International Relations Research," *European Journal of International Relations*, Vol. 28, no. 1 (2022), pp. 83-109.

[97] Cf. Michael Horowitz et al., "What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting," *Journal of Politics* 81, no. 4 (2019), pp. 1388-1404.

[98] Kertzer et al., "Hawkish Biases and Group Decision Making."

[99] See, for example, Janis, *Victims of Groupthink*. That may help to explain why this study's findings differ from Mandel and Barnes' analysis of underconfidence among Canadian intelligence professionals: even if most national security officials are intuitively prone to overconfidence, the organizational context in which they work may help to check – and, indeed, to overcorrect – that bias.

[100] Wayne et al., "Diplomacy by Committee."

[101] For example, the logic described above suggests that collaborating with cognitively diverse groups will typically mitigate overconfidence. We might thus expect these groups' performance to improve if they are given more time to analyze a decision. By contrast, we might expect that collaborating with cognitively homogeneous groups will typically exacerbate overconfidence such that these groups' performance will deteriorate if they are given more time to analyze a decision. Even simple questions such as "do people make better decisions when they have more time to conduct their analyses" are thus liable to be contingent on group structure, and the effects of group structure are, in turn, liable to be conditioned by other factors such as the time they have available to collaborate.

[102] Richard Clark, Lindsay R. Dolan, and Tyler Jost, "Bureaucratic Influence in International Politics," *Annual Review of Political Science*, forthcoming.

Apply the same corrections to advice you receive from others—the world is likely more uncertain than they think, too.

Second: remember that your judgments are prone to false positives. As described earlier, you can combat this problem by employing a "falsificationist" mindset and by avoiding "single-outcome forecasting." Instead of assessing the chances that a single statement is true, try to consider how uncertainty is distributed across multiple possibilities. Making this range of possibilities explicit can combat your natural tendency to fixate on one potential outcome to the exclusion of others.[103]

Finally, national security bureaucracies would benefit from providing personnel with quantitative feedback regarding their ability to assess uncertainty. Though this study documents widespread overconfidence among national security officials, its data also suggest that this bias is tractable. If just two minutes of feedback can substantially mitigate national security officials' overconfidence, then national security bureaucracies can almost certainly identify and combat cognitive biases at large scales. The procedure described in this article—administering surveys, processing data, and providing participants with individualized feedback on their performance—took roughly twenty minutes per cohort. These tasks can be automated, using code that appears in this article's replication materials. Similar exercises could be incorporated into any professional training program, or conducted by any institution willing to devote a small amount of time to the goal of improving its participants' judgments about the world.[104]

[103] Friedman and Zeckhauser, "Assessing Uncertainty in Intelligence," pp. 829-834.
[104] The Appendix contains instructions for replicating this study's procedures.