Vague Probability Assessments in National Security Decision Making: Exploring the Justifications for a Remarkably Uncontroversial Practice

Jeffrey A. Friedman
Paper prepared for the Bruno Kessler Foundation, December 2015

The central theme of this volume is that judging military effectiveness requires evaluating tradeoffs. These evaluations depend on probability assessments. The higher the chances of obtaining desired outcomes, the greater the costs a rational actor should be willing to pay to pursue them. As Ian Hacking thus writes, “Probability… lies at the basis of all reasonable choices made by officials. No public decision, no risk analysis, no environmental impact, no military strategy can be conducted without decision theory couched in terms of probabilities.”

The idea that probability surrounds military strategy is central to theory and practice. Carl von Clausewitz wrote that “no other human activity is so continuously or universally bound up with chance.” The U.S. Marine Corps’ capstone doctrine cites on its opening pages that “War is intrinsically unpredictable. At best, we can hope to determine possibilities and probabilities.” Contemporary international relations scholarship on war onset and coercive bargaining largely revolves around understanding the way that rational actors should form and revise probability assessments about the chances that they can achieve their goals through the use of force.

Yet regardless of how crucial probability assessment may be to the logical foundations of national security decision making, public officials often address this issue in strikingly vague ways. In their overview of Vietnam War decision making, for example, Leslie Gelb and Richard Betts explain that U.S. officials rarely discussed the chances that the measures they proposed would succeed. Instead, “Debates revolved around how to do things better, and whether they could be done, not whether they were worth doing.” Gelb and Betts identified just one major report produced by the Pentagon prior to deploying combat forces which assessed the feasibility

Jeffrey A. Friedman (jeffrey.a.friedman@dartmouth.edu) is Assistant Professor of Government at Dartmouth College. 10,038 words.

of these forces achieving their intended goals.\textsuperscript{6} This report concluded only that “there appears to be no reason we cannot win if such is our will,” while declining to offer any “assessment of the assurance the U.S. can have of winning.”\textsuperscript{7}

Of course, vague probability assessments appear in many fields. Psychologists, decision theorists, and public policy analysts have developed substantial literatures showing how “probability neglect” influences disciplines such as medicine, regulation, and law.\textsuperscript{8} In most professions, analysts generally dislike communicating probabilistic judgments explicitly (if at all), even when crucial decisions depend on those judgments.\textsuperscript{9} Yet national security is notable not just for the stakes involved with decisions made under uncertainty, but also for how deliberately its practitioners assess this uncertainty in vague terms. As this chapter explains, vague probability assessments are supported by official guidelines for intelligence analysis and military planning. And this practice is remarkably uncontroversial. Even public officials who oppose military actions like escalation in Vietnam or the Iraq Surge rarely push those policies’ advocates to clarify their probabilistic judgments.

There are several reasons why vague probability assessments are problematic in national security decision making. Recent research demonstrates that there are tangible benefits to expressing probability assessments precisely, even when those assessments rely on subjective judgments. Quantifying probabilistic judgments promotes accountability,\textsuperscript{10} helps analysts to self-evaluate and improve,\textsuperscript{11} and avoids miscommunications that result from using qualitative language.\textsuperscript{12} Even when analysts define “words of estimative probability” with lexicons involving

\begin{itemize}
  \item \textsuperscript{12} On natural variations in how people evaluate qualitative probability assessments, see Ruth Beyth-Merom, “How Probable is Probable? A Numerical Translation of Verbal Probability Expressions,” \textit{Journal of Forecasting}, Vol. 1
specific numeric ranges, readers still regularly interpret those terms in ways that their authors did not intend.\textsuperscript{13} Coarsening numeric probability assessments by translating them into qualitative terms consistently sacrifices predictive accuracy in foreign policy analysis.\textsuperscript{14} When national security officials choose to express probability assessments vaguely, they sacrifice all of these potential benefits. And if national security decision makers put lives and resources at risk in taking actions that logically rely on probabilistic judgments, then why would they \textit{not} scrutinize those judgments as carefully as possible?

As this chapter explains, there are three common justifications for vague probability assessments in national security: that there are no valid grounds for probabilistic precision in a field as complex and subjective as national security analysis; that even if probabilistic precision is justified in principle, it could have harmful behavioral consequences in practice; and that guidelines for expressing probability assessments would be impossible to implement over bureaucratic, political, or cultural obstacles. This chapter evaluates the credibility of these arguments in light of available scholarship in security studies and the decision sciences.

This review produces two main conclusions. First, claims that probabilistic precision is either unwise or infeasible rest on weak foundations. Some prominent objections to this practice are conceptually flawed; others are contradicted by rigorous research; and elsewhere there is little systematic evidence suggesting that probabilistic precision would harm the quality of national security decision making. There is little doubt that most national security analysts feel awkward or uncomfortable expressing subjective judgments precisely. But to say that some practice feels awkward does not mean that it is also a bad idea, and there is little credible evidence supporting the latter position. At the very least, this review encourages skepticism towards common defenses for leaving probability assessments vague when evaluating military strategies or when making national security decisions under uncertainty.

Second, although vague probability assessments in national security decision making are often defended on theoretical grounds – particularly with the idea that expressing subjective judgments precisely is conceptually inappropriate – the most important questions surrounding this practice are empirical in nature. If expressing probabilistic judgments precisely harms national security decision making, then this should have observable, empirical consequences. If not, then it is hard to understand why national security officials should sacrifice the readily available benefits of probabilistic precision, not least of which is simply removing unnecessary miscommunication from the decision making process. Given that probability assessments surround every intelligence estimate, every military plan, and every national security decision, even minor improvements in the way that national security officials address uncertainty could bring major aggregate gains. This is an area where scholars can play an important role informing the practice


\textsuperscript{14} Jeffrey A. Friedman, Joshua D. Baker, Barbara A. Mellers, Philip E. Tetlock, and Richard Zeckhauser, “The Value of Precision in Geopolitical Forecasting: Why Quantitative Probability Assessments Are Empirically Justifiable in Foreign Policy Analysis,” paper presented to the 111\textsuperscript{th} annual meeting of the American Political Science Association (San Francisco, Calif.: September 2015).
of national security analysis, while addressing questions that are central to broader research programs on high-stakes decision making.

The remainder of this chapter proceeds in five sections. Section 1 presents more information about vagueness in probability assessment, focusing especially on how official guidelines for intelligence analysis and military planning discourage practitioners from communicating probabilistic judgments precisely. Sections 2 through 4 scrutinize each of the three main justifications for vague probability assessments: that probabilistic precision is conceptually inappropriate, that it would have harmful behavioral consequences, and that it could not be implemented in practice. Section 5 concludes.

The examples in this chapter mainly relate to United States decision making, and specifically to the ways in which U.S. officials assess their chances of success in war. The chapter’s focus on the probability of success in war supports the book’s overall conception of military effectiveness as the degree to which militaries can accomplish desired outcomes at acceptable cost. This definition of military effectiveness emphasizes how one of the most fundamental challenges of military decision making is judging tradeoffs between the resources that combatants invest in their strategies and the chances that those strategies will succeed. \(^{15}\) Evaluating this balance is an extraordinarily difficult task which, I argue, is only complicated further when national security officials leave fundamental assumptions vague.

Yet the subject of probability assessment in war applies to virtually all national security decisions, and to all levels of war. When evaluating drone strikes, hostage rescue efforts, and special forces missions, for instance, military decision makers must assess the chances that they have correctly identified their targets, the risks of collateral damage, and potential harm to U.S. soldiers. Diplomatic negotiations require speculating about other states’ bargaining positions; managing procurement programs requires predicting future threat environments; financing national security programs requires projecting political and economic conditions several years down the line. All of these analyses involve confronting uncertainty. The question is thus not whether to assess probability in national security decision making but how it is most appropriate to do so. As this chapter will explain, this topic is surrounded by long-standing controversies and open empirical questions.

**Section 1. Vague Probability Assessments in National Security Decision Making: Common, Intentional, and Problematic**

In his 2009 “Commander’s Assessment” to President Obama, General Stanley McChrystal argued that the war effort in Afghanistan was failing, and that the United States should deploy 40,000 additional troops. General McChrystal justified this proposal by stating that it would “improve effectiveness,” and that it presented “the best prospect for success in this important mission.” Yet to say that a proposal offers the best chances of success does not mean that those chances are worth accepting. And in order to say whether the benefits of the Afghan Surge justified the costs, the key question was not _whether_ the Surge would improve effectiveness, but by _how much_ it would raise the chances of success. General McChrystal’s report did not address these questions:

\(^{15}\) Or the chances that those strategies can avoid unacceptable outcomes, such as nuclear war or financial insolvency.
this issue, even though it was logically crucial to evaluating the tradeoffs that his proposal presented.16

Similarly vague probability assessments surround many high-level national security decisions. In 1961, for instance, the Joint Chiefs of Staff were skeptical that a U.S.-backed invasion at the Bay of Pigs would successfully depose Fidel Castro. In internal discussions, the Joint Chiefs agreed that this plan’s chances of succeeding were roughly 3-in-10. Yet when the Joint Chiefs conveyed these views to President John F. Kennedy in writing, they stated only that “This plan has a fair chance of success.” President Kennedy apparently misinterpreted this statement to indicate optimism. When the invasion collapsed, the report’s author, Brigadier General David Gray, believed that his imprecise language had enabled a major strategic blunder.17 Kennedy, for his part, blamed the Chiefs for failing to provide clearer warning, complaining to a colleague that “Those sons of bitches with all the fruit salad just sat there nodding, saying it would work.”18 The resulting rift between the White House and the Joint Chiefs continued to grow throughout the early stages of the Vietnam War.19

---

18 Richard Reeves, President Kennedy: Profile of Power (New York: Simon and Schuster, 1993), p. 103
Figure 1. Guidelines for Expressing “Words of Estimative Probability” in the U.S. Intelligence Community


What We Mean When We Say: An Explanation of Estimative Language

When we use words such as “we judge” or “we assess”—terms we use synonymously—as well as “we estimate,” “likely” or “indicate,” we are trying to convey an analytical assessment or judgment. These assessments, which are based on incomplete or at times fragmentary information are not a fact, proof, or knowledge. Some analytical judgments are based directly on collected information; others rest on previous judgments, which serve as building blocks. In either type of judgment, we do not have “evidence” that shows something to be a fact or that definitively links two items or issues.

Intelligence judgments pertaining to likelihood are intended to reflect the Community’s sense of the probability of a development or event. Assigning precise numerical ratings to such judgments would imply more rigor than we intend. The chart below provides a rough idea of the relationship of terms to each other.

<table>
<thead>
<tr>
<th>Remote</th>
<th>Unlikely</th>
<th>Even chance</th>
<th>Probably, Likely</th>
<th>Almost certainly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. “Words of Estimative Probability” guidelines appearing in 2007 NIE, “Iran: Nuclear Intentions and Capabilities”

<table>
<thead>
<tr>
<th>Remote</th>
<th>Very unlikely</th>
<th>Unlikely</th>
<th>Even chance</th>
<th>Probably/ Likely</th>
<th>Very likely</th>
<th>Almost certainly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. “Words of Estimative Probability” guidelines recommended by the U.S. Director of National Intelligence in 2015

(a) For expressions of likelihood or probability, an analytic product must use one of the following sets of terms:

<table>
<thead>
<tr>
<th>almost no chance</th>
<th>very unlikely</th>
<th>unlikely</th>
<th>roughly even chance</th>
<th>likely</th>
<th>very likely</th>
<th>almost certain(ly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>remote</td>
<td>highly improbable</td>
<td>improbable (improbably)</td>
<td>roughly even odds</td>
<td>probable (probably)</td>
<td>highly probable</td>
<td>nearly certain</td>
</tr>
<tr>
<td>01-05%</td>
<td>05-20%</td>
<td>20-45%</td>
<td>45-55%</td>
<td>55-80%</td>
<td>80-95%</td>
<td>95-99%</td>
</tr>
</tbody>
</table>
The Bay of Pigs episode is notable for how the Joint Chiefs intentionally left their probability assessment imprecise. Translating a quantitative probability estimate into qualitative language may have seemed more natural or appropriate in the context of their report, yet this caused needless miscommunication that ultimately left both sides worse off. In principle, one would expect that national security officials would take care to avoid such miscommunication. Yet there are many areas of national security analysis where public officials are in fact instructed to express probabilistic judgments vaguely.

Figure 1, for example, presents three “words of estimative probability” spectrums which the U.S. Intelligence Community (IC) has employed since 2007. The IC developed these guidelines in response to legislation mandating that intelligence analysts “properly caveat and express uncertainties or confidence in analytic judgments.” Each of these guidelines instructs analysts to communicate probability in qualitative terms. The most recent version of these guidelines, contained in the U.S. Director of National Intelligence’s (DNI’s) directive on “Analytic Standards,” defines these terms in relation to specific numeric ranges. In order to apply this guidance properly, analysts must determine where a probability estimate falls along the number line, then coarsen that estimate using the appropriate qualitative term. By defining a lexicon for interpreting such terms, the DNI’s analytic standards prevent extreme miscommunication as in the Bay of Pigs episode, but the difference is degree rather than kind: analysts are still instructed to make their probability assessments intentionally imprecise.

Though “words of estimative probability” spectrums are most often debated in intelligence studies, similar guidelines appear in other areas of national security. For example, Figure 2 provides a risk assessment matrix from current U.S. Army doctrine. According to this doctrine, planners are expected to assess risks surrounding proposed operations using five qualitative terms: “frequent,” “likely,” “occasional,” “seldom,” and “unlikely.” These terms have confusing definitions. The word “frequent” is defined as “Occurs very often, known to happen regularly. In illustration, given 500 exposures to the hazard, expect that it will definitely happen to someone.” The word “likely” is defined as something that “will occur at some point” in 1,000 exposures.

Three aspects of these definitions are worth noting. First, these are very different from intuitive definitions of words like “frequent” or “likely,” which usually denote probabilities over fifty percent. Second, the notion that analysts should “expect” something to “definitely happen” is incoherent. Third, these definitions involve an oddly ambiguous use of numeric definitions. Does “frequent” mean 1-in-500? Is it more like 1-in-700 (roughly the chances needed for an

---

20 IRTPA Section 1019(b)(2)(A).
24 The two sentences of the manual’s definition of the word “frequent” seem especially nonintuitive, equating the phrase “very often” with an odds ratio of perhaps 1-in-500.
25 Similarly, the only way to guarantee that a hazard “will occur at some point” in 1,000 exposures is to assume that its probability is 100 percent.
Figure 2. U.S. Army guidelines for communicating risk


<table>
<thead>
<tr>
<th>Severity</th>
<th>Hazard Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catastrophic</td>
<td>E</td>
</tr>
<tr>
<td>Critical</td>
<td>E</td>
</tr>
<tr>
<td>Marginal</td>
<td>H</td>
</tr>
<tr>
<td>Negligible</td>
<td>M</td>
</tr>
</tbody>
</table>

HAZARD PROBABILITY (The likelihood that an event will occur)
- Frequent - Occurs often, continuously experienced.
- Likely - Occurs several times.
- Occasional - Occurs sporadically.
- Seldom - Unlikely, but could occur at some time.
- Unlikely - Can assume it will not occur.

B. Definitions of probabilistic terms from U.S. Army Field Manual 5-19, “Composite Risk Management”

Assess Each Hazard on the Probability of the Event or Occurrence

1-23. Probability is the likelihood of an event. This is your estimate, given what information you know and what others have experienced. The probability levels estimated for each hazard are based on the mission, COA, or frequency of a similar event. For the purpose of CRM, there are five levels of probability—frequent, likely, occasional, seldom, and unlikely:

- Frequent – Occurs very often, known to happen regularly. In illustration, given 500 or so exposures to the hazard, expect that it will definitely happen to someone. Examples of frequent occurrences are vehicle rollovers, rear-end collisions, and heat injury during a battalion physical training run with hot weather or nonacclimated Soldiers.
- Likely – Occurs several times, a common occurrence. In illustration, given 1000 or so exposures without proper controls, it will occur at some point. Examples might include improvised explosive devices (IEDs), wire strikes for aircraft, controlled flights into terrain, and unintentional weapons discharges.
- Occasional – Occurs sporadically, but is not uncommon. You may or may not get through your deployment without it happening. Some examples might include unexploded ordnance (UXO) and fratricide.
- Seldom – Remotely possible, could occur at some time. Usually several things must go wrong for it to happen. Examples might include things like heat-related death or electrocution.
- Unlikely – Can assume will not occur, but not impossible. Examples might include detonation of containerized ammunition during transport.
outcome to be likely in at least one of 500 independent trials)? Or is it 1-in-100 (roughly the chances needed for a 99 percent probability of occurrence in 500 independent trials)? While the Army’s risk assessment doctrine ostensibly attempts to clarify the use of probabilistic language, the resulting guidelines leave substantial space for interpretation.

Figure 3 presents current guidelines for communicating probability when assessing the reliability of human intelligence sources. Whenever national security analysts consider information from a human source, it is vital to assess the chances that this source could be lying. One of the principal critiques of the 2003 National Intelligence Estimate on Iraq’s weapons of mass destruction programs, for instance, is that the IC did not adequately convey skepticism about the validity of information provided by Iraqi defectors such as the infamous “Curveball.”

Human intelligence reports are thus generally accompanied by source reliability assessments.

![Figure 3. U.S. Army guidelines for communicating source reliability](image)


In order to convey those assessments consistently, the U.S. military encourages analysts to use six qualitative terms: “reliable,” “usually reliable,” “fairly reliable,” “not usually reliable,” “unreliable,” and “cannot be judged.” These qualitative statements of probability are, in turn, defined in relation to other qualitative statements of probability, such as “no doubt,” “minor doubt,” “doubt,” and “significant doubt.” There is no indication as to how borders between these terms should be drawn. If there is a ten percent chance that a source is lying, this could conceivably fall under any of the six terms presented in the doctrine.

---

26 Senate Select Committee on Intelligence, Report on the U.S. Intelligence Community’s Prewar Intelligence Assessments on Iraq (2004), ch. IV.
In principle, the guidelines presented in Figures 1 through 3 apply, respectively, to every key judgment in every intelligence report, every risk assessment in every Army plan, and every source assessment provided by the U.S. military. Many consequential judgments and decisions depend on these sources. And though one might expect that national security officials would take special care to assess the probability of higher-profile issues like the chances that military operations will achieve their intended goals, this is not always the case. For example, U.S. Army Field Manual 5-0, The Operations Process, states that when commanders evaluate potential courses of action (COAs),

The staff compares feasible COAs to identify the one with the highest probability of success…. The selected COA should also pose the minimum risk to the force and mission accomplishment; place the force in the best posture for future operations; provide maximum latitude for initiative by subordinates; provide the most flexibility to meet unexpected threats and opportunities; provide the most secure and stable environment for civilians in the AO; [and] best facilitate initial information themes and messages.28

These relational judgments are important, as decision makers should always seek to maximize their chances of success and to minimize risks. But it is not sufficient to identify whether a course of action has the “highest probability of success” or whether it poses “the minimum risk to the force.” Evaluating cost-benefit tradeoffs also requires estimating chances and risks in absolute terms. In this respect, General McChrystal’s statement about how the Afghan Surge offered the “best prospect for success” left logical gaps for decision making, but it was also consistent with official guidelines for evaluating military actions.

Another problematic way to address uncertainty in military planning is to analyze what measures are necessary for success without also assessing the chances that those measures will also be sufficient. For example, U.S. troop surges in both Iraq and Afghanistan were justified, in part, based on force sizing guidance from U.S. Army Field Manual 3-24, on counterinsurgency (COIN). The relevant paragraph of this doctrine explains that “Twenty counterinsurgents per 1000 residents is often considered the minimum troop density required for effective COIN operations; however as with any fixed ratio, such calculations remain very dependent upon the situation.”29 At a glance, this statement appears to offer a clear basis for planning: if commanders wish to wage effective counterinsurgency, they should send at least twenty troops per thousand residents in an area of operations.

Yet the doctrine provides no basis for inferring what the chances of success might be if this threshold is met. Note that the “twenty troops per thousand inhabitants” rule of thumb is presented as the “minimum troop density required for effective COIN operations.” This implies that commanders are likely to fail if they fall below the threshold, but it says nothing about the chances of succeeding if planners meet this requirement. Even this limited claim is preceded in Field Manual 3-24 with the caveat that “such calculations remain very dependent upon the situation.” Taken literally, the guidance only suggests that if commanders surpass the stated troop density threshold, then it is possible that success will then be possible. Again we see doctrine explicitly acknowledging that crucial decisions reside in the realm of chance, but declining to address those chances directly.

Of course, there is a simple fix for the problems that these guidelines raise: when national security analysts assess probability, they could express their judgments quantitatively, using numeric percentages (e.g., “10 percent”), odds ratios (e.g., “20-to-1”) or better’s odds (e.g., “1-in-5”). All of the confusion described in this section – how to evaluate the Afghan Surge’s prospects, how to communicate skepticism about the Bay of Pigs, what it means to say that a risk is “likely” or that a source is “fairly reliable” – would be dispelled if analysts quantified their assessments of probability.

There are several reasons to be concerned with the fact that national security analysts do not generally do this. Vague probability assessments can bias decision makers. For example, pharmaceutical companies routinely goad customers into purchasing treatments that promise to reduce risks, even if those risks are already vanishingly small. Even if national security officials do not seek to fool their colleagues in this manner, it is reasonable to expect a similar result: that decision makers would find it easier to accept a strategy billed as the “best chance of success,” or one which will “improve effectiveness,” over a strategy whose odds of succeeding are “roughly one in five,” even if those strategies are the same. Similarly, warning decision makers that they will fail without implementing a policy surely provides more encouragement for taking action than stating that the policy has a small chance of success, even though both statements can simultaneously be true.

Leaving probability assessments vague can also enable tendencies to avoid carefully analyzing the risks and uncertainties surrounding costly actions. International relations scholars generally argue that national security decision makers are inclined towards overoptimism, that they tend to focus on the desirability of long-term goals rather than the feasibility of achieving those goals, and that their decisions often rely on intuitive judgments rather than rigorous assessments of tradeoffs. Even if probability assessment is logically crucial to logical decision making, this is no guarantee that national security officials will naturally pay close attention to this subject. When presented with ambiguous assessments, decision makers have political incentives and natural inclinations to interpret those assessments in self-serving ways.

Thirty years after the Johnson administration sent U.S. combat troops to Vietnam, then-Secretary of Defense Robert McNamara reflected on the strange absence of internal discussions

about the United States’ chances of success in this conflict. McNamara explained that “we never carefully debated what U.S. force would ultimately be required, what our chances of success would be, or what the political, military, financial, and human costs would be if we provided it. Indeed, these basic questions went unexamined. We were at the beginning of a slide down a tragic and slippery slope.” McNamara admitted that “It seems beyond understanding, incredible, that we did not force ourselves to confront such issues head-on.”

The fact that national security officials do not automatically address these issues carefully is exactly the reason to build explicit probability assessment into official guidelines for intelligence analysis and military planning. At the very least, the fact that official guidelines actively encourage vague probability assessment deserves scrutiny. The remainder of this chapter examines three prominent justifications for this practice.

Section 2. Can Probabilistic Precision Be Justified?

In February 1965, National Security Advisor McGeorge Bundy wrote President Lyndon Johnson to warn that “The situation in Vietnam is deteriorating, and without new U.S. action defeat appears inevitable.” Bundy explained that escalating the war through strategic bombing and the deployment of combat forces provided “the best available way of increasing our chance of success.” But Bundy acknowledged that he had little sense of just how helpful these measures would be, writing that “We cannot estimate the odds of success with any accuracy – they may be somewhere between 25% and 75%.”

Bundy’s vague estimate of the chances that escalation would succeed in Vietnam stands in contrast to the way that he and the Pentagon’s “Whiz Kids” generally sought to analyze and to quantify as much of the war effort as possible. As a former professor of political science at Harvard, where he was always Dean of the Faculty of Arts and Sciences, Bundy would have known that evaluating tradeoffs in high-stakes decision making requires rigorously assessing uncertainty. By comparison, imagine if Bundy had caveated his recommendation by saying that “We cannot estimate the troop levels this requires with any accuracy – they may be somewhere between 250,000 and 750,000.” This would have been an obviously insufficient basis for making a major decision. Yet from the standpoint of evaluating cost-benefit tradeoffs, vagueness about force requirements would have been no more problematic than declining to estimate the probability that those forces would succeed.

Nevertheless, many scholars and practitioners of national security would consider Bundy’s hesitation to offer a clear probability assessment to be entirely appropriate. A common argument within strategic studies is that war is an integratively complex, nonlinear process which surpasses

---

36 Logically speaking, vague troop levels would actually be less problematic. The expected value of any decision is its expected benefits minus costs: \( pB - C \), where \( p \) is the probability of success, \( B \) is the benefit of a successful outcome and \( C \) is the decision’s cost. Troop levels are just one element of a military strategy’s costs, whereas a strategy’s expected benefits are directly proportional to its perceived probability of success.
the limits of any known analytic techniques. The crux of this argument is that assessing military outcomes is an inherently subjective process that cannot sustain precise conclusions. This is perhaps the most commonly-cited justification of vague probability assessments in national security decision making: even if this vagueness has downsides, many people believe it is simply inappropriate to do anything different.

There are three reasons why this argument does not withstand scrutiny. First, any probabilistic belief, no matter how subjective, can be validly translated into a single numeric assessment. To demonstrate this point, decision theorists often invoke a thought experiment involving the comparison of lotteries. For example, consider what you believe are the chances that a Republican candidate will win the next U.S. presidential election. Now consider which of two gambles you prefer. In Gamble A, you are paid $1,000 if the Republican candidate wins the next presidential election. (If the Republican candidate does not win, you receive nothing). In Gamble B, we flip a coin on election night, and if it comes up heads then you win the $1,000. You can only take one of these bets. Would you prefer the coin flip? If so, this implies that you believe the chances of a Republican win are no greater than 50 percent. In principle, we can toggle the probabilities in Gamble B until you are indifferent between choices. This would represent your subjective probability estimate that a Republican will win the next U.S. presidential election.

Comparing lotteries is one of several methods for eliciting subjective probabilities, and the point is not to suggest that national security officials should actually gamble money on their probability assessments. The takeaway is simply that when analysts assess probability in national security of any other context, there are valid ways to describe such views using numbers, subjective foundations and all. The notion that it is conceptually impossible to quantify these beliefs is simply incorrect. The key move is to understand that the validity of these estimates derives not from objective properties of empirical data or from mathematical deduction, but rather from identifying conclusions that are consistent with analysts’ beliefs. If an analyst then

---


38 Another argument is that


42 The origins of this “subjectivist” approach to probability, developed by Frank Ramsey and Bruno de Finetti, is described in Donald Gillies, Philosophical Theories of Probability (New York: Routledge, 2000). As Frank Lad puts it: “In the proper syntax of the subjectivist formulation, you might well ask me and I might well ask you, ‘What is
believes it is important to qualify her degree of confidence in making such an estimate than this can convey useful, additional information. But likelihood and confidence are different concepts, and regardless of how reliable an analyst believes a probability assessment to be, she can always express that estimate using a single number.\textsuperscript{43}

A second reason why conceptual opposition to probabilistic precision is unsustainable is that if quantitative probability assessments are incoherent, then one cannot accept quantitative judgments that are based on probability assessments either. For example, consider General McChrystal’s claim that 40,000 additional forces in Afghanistan represented “the minimum force levels to accomplish the mission with an acceptable level of risk.”\textsuperscript{44} This statement implied that a smaller troop increase (say, 20,000 forces) would produce meaningfully smaller chances of success. If McChrystal could not make valid probability assessments about strategic outcomes, then his recommendation would have been meaningless. And if quantitative precision was acceptable for conveying McChrystal’s force estimates, then it must also have been acceptable for expressing the assumptions behind those estimates: conclusions can be no more justifiable than premises they are based on.

Of course, no one believes that it is inappropriate for national security officials to quantify the number of troops, money, or time that a military operation is expected to require. These estimates are inherently subjective and they require assessing uncertainty, but military decision makers cannot do their jobs without rendering these estimates into quantitative form. It is impossible to send “a fair number” of troops overseas or to authorize “significant” funding to support them; at some point, specific numbers must be entered into orders and legislation. It is possible to go to war without estimating the chances of success in a meaningful way, but this does not mean that it is any less appropriate to address this aspect of decision making explicitly.\textsuperscript{45}

\textit{your probability for a specified event?} It is proposed that there is a distinct (and generally different) correct answer to this question for each person who responds to it…. Your answer can be evaluated as correct or incorrect only in terms of whether or not you answer honestly.” Lad, \textit{Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction} (New York: Wiley, 1996), pp. 8-9.

\textsuperscript{43} In principle, probability refers to the chances that a statement is true, and confidence qualifies the reliability of that estimate. For example, you would presumably say that the probability of a coin flip coming up heads is 50 percent, and you are presumably highly confident in that estimate. Yet if asked to estimate the probability that a Republican candidate will win the next U.S. presidential election, you might say 50 percent with low confidence, because the topic is so difficult to judge. Combining estimates of likelihood and confidence (e.g., by providing a confidence interval without a point estimate) conflates two distinct concepts. For more on the distinction between likelihood and confidence, and how this distinction is often not observed in national security analysis, see Friedman and Zeckhauser, “Assessing Uncertainty in Intelligence,” and idem, “Handling and Mishandling Estimative Probability.”

\textsuperscript{44} COMISAF, \textit{Initial Assessment}.

\textsuperscript{45} To restate the point, if national security analysts cannot make meaningful probability assessments, then there would never be a valid reason to expend substantial resources on a military strategy: the cheapest option would always be the most defensible, because no one could reasonably claim that its chances of success are lower than the alternatives. Rejecting this view requires believing that probability assessments in national security have some degree of validity. And if these assessments have some degree of validity, then it is problematic to leave them intentionally vague. For a related critique of extreme views on nonlinearity, see Richard K. Betts, “Is Strategy an Illusion?” \textit{International Security}, Vol. 25, No. 2 (Autumn, 2000), p. 20. See also Robert A. Pape, “The Air Force Strikes Back,” \textit{Security Studies}, Vol. 7, No. 2 (Winter 1997/98), pp. 196-97.
One might argue that even if these probabilistic judgments have some validity, probability assessment is simply so subjective that anything beyond a rough degree of precision is essentially random noise. If existing “words of estimative probability” spectrums capture all of the information that analysts can provide, then why go further than this? One answer is that intuitive definitions of “words of estimative probability” vary substantially, both across individuals and between contexts. These intuitive definitions are difficult to overcome. For example, recent research demonstrates that even when respondents are presented with lexicons translating verbal expressions into probabilistic ranges, they often still interpret “words of estimative probability” in ways that authors did not intend. In this respect, quantitative precision is justifiable simply as a method of eliminating miscommunication.

But another reason why such opposition to quantitative probability assessment is unjustified is that foreign policy analysts actually can effectively parse probabilities with numeric precision. A recent study examined more than 800,000 predictions collected by the Good Judgment Project, a research program sponsored by the U.S. Intelligence Community. Over a four-year period, Good Judgment Project participants offered numeric probability estimates of potential geopolitical events. The authors of this study rounded those estimates into “words of estimative probability” consistent with official guidelines described in the previous section. This rounding resulted in a statistically significant loss of predictive accuracy across a wide range of forecasters, time frames, and question types. Quantitative probability assessments are thus not just defensible in theory: there are empirical grounds for believing that they convey more information than what common qualitative expressions allow.

Section 3. Does Probabilistic Precision Have Harmful Behavioral Consequences?

The previous section argued that quantitative probability assessments are both theoretically coherent and empirically defensible, even in a field as complex and subjective as national security decision making. As a result, if scholars or public officials oppose probabilistic precision in national security analysis, they must base their arguments on the behavioral consequences of this practice. In other words, even if quantitative probability assessments are appropriate in principle, they could still generate negative side effects in practice. Generally speaking, these behavioral arguments fall into two categories. First, quantifying probability assessments could distort decision makers’ perceptions by making subjective judgments seem more rigorous than they really are. Second, if most analysts reason naturally in qualitative terms, they might argue that even if these probabilistic judgments have some validity, their subjective nature makes anything beyond a rough degree of precision essentially random noise.
then asking them to translate their beliefs into quantitative language could cause otherwise avoidable errors.\(^{50}\)

From a rationalist standpoint, neither argument is compelling. Logically speaking, rational decision makers should not be influenced by the ambiguity surrounding the assessments they receive.\(^{51}\) And analysts cannot use “words of estimative probability” lexicons properly without determining how their probabilistic judgments map onto the number line, or else they would not be able to choose the appropriate terms.

Nevertheless, scholarship on judgment and decision making is replete with examples of how individuals consistently respond to uncertainty in nonrational ways.\(^{52}\) The behavioral consequences of probabilistic precision must therefore be taken seriously. Indeed, this has been a long-standing subject for empirical research in the decision sciences.

Generally speaking, this research indicates that behavioral concerns with quantifying probability assessments are overstated. As one paper summarized, “The surprising result emerging from all these studies is the approximate equivalence in peoples’ ability to use verbal and numerical expressions of probability over a range of tasks and paradigms.”\(^{53}\) It appears that decision makers pay closer attention to probability assessments expressed in numerical form, but there is little evidence that quantification consistently biases the way that individuals interpret the meaning of these assessments.\(^{54}\)

Of course, these laboratory experiments, typically involving undergraduate respondents, cannot directly falsify expectations about how national security officials would behave in real situations.\(^{55}\) Yet if objections to probabilistic precision in national security decision making rely on claims about the ways in which quantifying subjective judgments would corrupt human

\(^{50}\) On these arguments, see the sources listed in note 34, along with Alf C. Zimmer, “A Model for the Interpretation of Verbal Predictions,” *International Journal of Man-Machine Studies*, Vol. 20 (1994), pp. 121-134. One version of this perspective holds that, since mathematical reasoning developed relatively late in human evolution, the human brain is conditioned to reason qualitatively.

\(^{51}\) Key to this argument is that risk preferences stem from the way decision makers evaluate outcomes and this should not influence the manner in which they evaluate probabilities. Reducing ambiguous beliefs about probabilities into single point estimates is therefore theoretically defensible in a manner that cannot be extended to dealing with uncertainty about a policy’s outcome. On the difference between rational and behavioral responses to probabilistic ambiguity, see Daniel Ellsberg, “Risk, Ambiguity, and the Savage Axioms,” *Quarterly Journal of Economics*, Vol. 75, No. 4 (1961), pp. 643-669. On how rational decision makers should not have risk preferences when evaluating probabilities, also see Savage, *Foundations of Statistics*.


\(^{55}\) A related question is why one should favor quantifying probability assessments if laboratory experiments find that this does not impact the quality of decision making. Section 5 returns to this point.
judgment, it should be possible to identify these harmful effects empirically. At present, there is little systematic evidence backing this view.

One difficulty in testing skeptics’ contentions directly is that opponents of probabilistic precision in national security decision making rarely translate their behavioral claims into falsifiable hypotheses. For example, imagine that instead of asserting that the Afghan Surge presented “the best prospect for success in this important mission,” General McChrystal had written that this proposal had a 35 percent chance of achieving its intended goals. Skeptics of probabilistic precision would argue that quantifying McChrystal’s assessment would then bias President Obama’s reaction, on the grounds that McChrystal’s judgment would have appeared to be more rigorous than it really was. But what would this bias entail? Would President Obama have been more likely to take a risky action on the grounds that his top commander had provided a seemingly-rigorous assessment? Or would the additional clarity have made the risk seem so great that President Obama would then be less willing to accept it? Though the notion that numeric probability assessments seem more rigorous than they really are is commonly stated by scholars and practitioners of national security, few authors articulate the observable implications of the point, let alone test those implications empirically.\footnote{For preliminary analysis of this subject, see Jeffrey A. Friedman, Jennifer S. Lerner, and Richard Zeckhauser, “Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Officials,” paper prepared for the 57th annual meeting of the International Studies Association (Atlanta, Ga: March, 2016).}

Similarly, if quantifying subjective judgments would harm the quality of assessments made by national security analysts, it is unclear what manifestation this problem would take. Would quantitative assessments make probability estimates more extreme (i.e., more overconfident)? Would analysts who are uncomfortable expressing subjective judgments precisely instead become underconfident because they are unwilling to “go out on a limb” explicitly? And even if the direction of this bias were confirmed, one cannot say whether that bias would entail harmful effects without forming assumptions as to whether national security analysts tend to be over- or underconfident to begin with. All of these questions remain wide open for empirical research.\footnote{The standard assumption is that national security analysts tend towards overconfidence: for example, see Richards J. Heuer, Jr., Psychology of Intelligence Analysis (Washington, D.C.: Center for the Study of Intelligence, 1999), as well as Philip E. Tetlock, Expert Political Judgment: How Good Is It? How Do We Know? (Princeton, N.J.: Princeton University Press, 2005). However, a recent, large-scale study of Canadian intelligence analyses found that probability assessments were consistently underconfidence: David R. Mandel and Alan Barnes, “Accuracy of Forecasts in Strategic Intelligence,” Proceedings of the National Academy of Sciences, in press. Note that Mandel and Barnes’s findings do not conflict with the research, cited above, on how national security decision makers tend towards overconfidence.}

What recent research does credibly indicate is that when foreign policy analysts receive feedback on quantitative assessments, they quickly learn to calibrate those assessments effectively. Though this is a long-established finding in the decision sciences,\footnote{For example, see Marc Alpert and Howard Raiffa, “A Progress Report on the Training of Probability Assessors” in Daniel Kahneman, Paul Slovic, and Amos Tversky eds., Judgment Under Uncertainty (New York: Cambridge University Press, 1982).} the Good Judgment Project’s research has demonstrated the benefits of training and feedback within the specific domain of foreign policy analysis, and on an unprecedented scale. Thus when Good Judgment Project respondents say that there is a 20 percent that some outcome will occur, we actually observe those outcomes occurring roughly 20 percent of the time. There are important, open questions about how easily this method of assessment and feedback could be implemented...
among national security professionals. However, the Good Judgment Project’s findings provide clear evidence countering common claims that foreign policy analysts should be unable to quantify their subjective assessments effectively.

Moreover, existing experience with quantifying estimative probabilities in intelligence analysis provides little reason to believe that this practice meaningfully biases analysts or decision makers. In recent articles, James Marchio described a U.S. Defense Intelligence Agency (DIA) experiment with quantifying probability assessments in the 1970s, while Alan Barnes reported on a similar, contemporary effort within the Canadian intelligence service. Neither author reported that quantifying probability assessments systematically biased analysts towards over- or under-confidence. Nor did these experiences suggest that quantifying probabilities corrupted decision makers’ abilities to evaluate intelligence reports effectively.

Instead, both studies found that the main difficulties with these experiments were convincing analysts to do something they found to be out of the ordinary. Analysts asked reasonable questions and conveyed natural skepticism about these experiments (and the DIA’s experiment did not survive the arrival of new managers who shared that skepticism). But there was little indication that quantifying subjective judgments led to serious, behavioral side effects. Similarly, Zvi Lanir and Daniel Kahneman reported that an Israeli experiment with quantifying probability assessments tapered off because decision makers seemed surprisingly uninterested in this information, not because that information seemed to have harmful consequences.

If national security officials find it awkward to discuss probability assessments precisely, then this is important to understand. But to say that some practice is uncomfortable does not mean that it is harmful, and these claims should not be confused. An important question which this discomfort raises, however, is whether probabilistic precision could feasibly be implemented in practice. This is the subject of the following section.

**Section 4. Can Probabilistic Precision Be Implemented?**

A third common justification for probabilistic vagueness is that there is really no feasible alternative. Even if probabilistic precision is justified in principle, and even if it would not entail negative behavioral consequences in practice, skeptics might claim that institutional resistance to this idea would be so strong that pursuing it is a fool’s errand.

Bureaucracy, politics, culture, and personal self-interest all influence national security decision making. One cynical argument is that national security analysts prefer to assess probabilities using “weasel words” so as to avoid criticism when their estimates appear to be mistaken after the fact. Such behavior would be consistent with a broader strategy of blame

---

59 Tetlock and Gardner, *Superforecasting*.  
62 See the concluding chapter of Tetlock and Gardner, *Superforecasting*, for a discussion of these issues.  
avoidance in public policy analysis that Christopher Hood calls “abstinence”: intentionally declining to provide information that could expose public officials to criticism after the fact. A different argument is that the culture of many national security analysts resembles “poets” as opposed to “mathematicians,” and that practitioners would strongly resist perceived changes to this culture whatever the merits of those changes might be.

Of course, these arguments do not mean that vague probability assessments are actually defensible, or that observers should not question them. And national security policy contains many ideas that initially met skepticism and bureaucratic resistance, only to prove surprisingly feasible (and successful) in retrospect. The creation of the all-volunteer army and the interservice restructuring accomplished by the Goldwater-Nichols Act are two prominent examples. Originally, these proposals were thought to be undesirable; later they were thought to be desirable but impossible to implement; today they are some of the most important foundations of U.S. national security.

Moreover, there are two specific reasons to believe that institutional resistance to probabilistic precision in national security is not as indomitable as some people expect. First, the U.S. Intelligence Community’s (IC’s) analytic standards have undergone sweeping changes in the last decade that previously seemed unthinkable. The original idea of using “words of estimative probability” in intelligence – that is, defining qualitative terms in relation to specific numeric ranges – was originally proposed by Sherman Kent in 1964. Figure 4 presents the “Odds Table” that Kent recommended based on surveys of how national security officials intuitively defined different phrases. For decades, Kent’s idea gained little traction, leading to the perception that the IC’s “poets” presented an insurmountable obstacle to institutional change. As of this writing, the Central Intelligence Agency’s (CIA’s) website still prefaces the online

<table>
<thead>
<tr>
<th>The General Area of Possibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Certainty</td>
</tr>
<tr>
<td>93% give or take about 6%</td>
</tr>
<tr>
<td>75% give or take about 12%</td>
</tr>
<tr>
<td>50% give or take about 10%</td>
</tr>
<tr>
<td>30% give or take about 10%</td>
</tr>
<tr>
<td>7% give or take about 5%</td>
</tr>
<tr>
<td>0% Impossibility</td>
</tr>
</tbody>
</table>

Figure 4. Sherman Kent’s 1964 “Odds Table”

Moreover, there are two specific reasons to believe that institutional resistance to probabilistic precision in national security is not as indomitable as some people expect. First, the U.S. Intelligence Community’s (IC’s) analytic standards have undergone sweeping changes in the last decade that previously seemed unthinkable. The original idea of using “words of estimative probability” in intelligence – that is, defining qualitative terms in relation to specific numeric ranges – was originally proposed by Sherman Kent in 1964. Figure 4 presents the “Odds Table” that Kent recommended based on surveys of how national security officials intuitively defined different phrases. For decades, Kent’s idea gained little traction, leading to the perception that the IC’s “poets” presented an insurmountable obstacle to institutional change. As of this writing, the Central Intelligence Agency’s (CIA’s) website still prefaces the online

---

65 Johnston, *Analytic Culture in the U.S. Intelligence Community*.
67 Kent, “Words of Estimative Probability.” A classified version of this article appeared even earlier.
version of Kent’s article by stating that “Although Sherman Kent's efforts to quantify what were essentially qualitative judgments did not prevail, the essay's general theme remains important today.”

Yet Kent’s idea did ultimately prevail. As we saw in Section 1, the U.S. Director of National Intelligence’s (DNI’s) current analytic standards not only seek “to quantify what are essentially qualitative judgments” (as the CIA’s website puts it), but do this even more explicitly than what Kent had originally recommended, as the DNI’s current “words of estimative probability” spectrum contains seven terms instead of Kent’s five. Thus while obstacles to probabilistic precision are important to acknowledge, one should not overstate the challenges they pose to institutional change. A prominent intelligence scholar calls the IC’s progress on this front “revolutionary.” One might frame the quantification of subjective probability estimates as the next logical step in this ongoing revolution.

Moreover, the quantification of probability assessments does not require revising institutional guidelines if decision makers are willing to prompt their advisers to provide clearer information. During the search for Osama bin Laden in 2010-11, for instance, the CIA initially reported to President Obama that there was a “serious possibility” that Al Qa’ida’s leader was living in a compound in Abbottabad, Pakistan. When President Obama pushed analysts to describe what a “serious possibility” meant in clearer terms, they offered numeric assessments. For example, one CIA official put the chances of bin Laden being in Abbottabad at 95 percent, while the Agency’s Deputy Director Michael Morell offered 60 percent. Morell explained this disparity to President Obama by arguing that while many counterterrorism officials had understandably grown confident in their targeting abilities, his experience assessing Iraq’s weapons of mass destruction programs a decade earlier left him wary of drawing conclusions from circumstantial evidence.

In this way, unpacking a vague estimate (“a serious possibility”) revealed substantive disagreements and raised fundamental issues about analytic and cognitive constraints on forming judgments under uncertainty. More importantly for our purposes here, analysts did not insubordinately refuse to state subjective judgments precisely when asked to do so. Ultimately, if a decision maker requests a particular kind of information, one should expect that analysts will provide it.

Of course, one might argue that many national security decision makers prefer estimative imprecision in order to avoid ex post criticism of their actions. U.S. Secretary of State Dean Acheson famously argued that public officials need to be able to make controversial policies “clearer than truth” in order to sell them to the public (and perhaps, to believe in those policies themselves). Indeed, it is difficult to explain why public officials like Lyndon Johnson, Robert McNamara, and McGeorge Bundy failed to assess their chances of success in Vietnam more directly without accounting for some degree of defensive avoidance.

---

71 On political incentives to avoid intelligence analysis, see Rovner, Fixing the Facts.
Yet this is where multiple advocacy can play an important role insuring that decision makers do not neglect to consider the foundations of their decisions explicitly. When Bundy said that the chances of success in Vietnam might be “somewhere between 25% and 75%,” this provided a clear opening for skeptics to question how he could possibly support a military strategy without a better understanding of its prospects. Forming such an assessment would no doubt be difficult, but without being able to defend these assumptions, a public official also cannot logically justify putting lives and resources at risk. Similarly, when General David Petraeus testified to Congress in 2007 supporting the Iraq Surge, he famously described this strategy as being “hard, but not hopeless.”73 This phrase provided no reason to think that the Surge’s prospects were good enough to accept. If anything, Petraeus’s statement indicated that the Surge was exactly the kind of low-probability, high-consequence gamble where it is especially important for decision makers to assess uncertainty as clearly as possible.

As a general defending his strategy in a public forum, it is unsurprising that Petraeus would not volunteer information about the long odds that he faced: according to later reports, some members of his staff may have believed that these odds were as low as 10 percent.74 What is more notable about this experience is that no members of Congress – even those who vehemently opposed the Surge – seized this opportunity to press their witness for more detail on what a “hard, but not hopeless” probability entailed. If Petraeus made clear that he thought the probability of success was small, critics could have used this information to cudgel his testimony. If Petraeus had been unable to clarify his beliefs, then critics could have reasonably asked how he could logically support a military strategy without estimating its chances of success.

In this respect, the lack of probabilistic precision in national security decision making cannot simply be attributed to political or bureaucratic motives. Even if some actors have interests in keeping their probability assessments vague, other actors have incentives to press the issue. This is a place where multiple advocacy should, in principle, force national security officials to explore the logical foundations of their choices in close detail. The fact that even the political opponents of controversial national security decisions do not insist on eliciting this information reflects the remarkably uncontroversial nature of vague probability assessments in national security decision making.

Section 5. Studying Probability Assessment in National Security Decision Making: Moving from Epistemology to Empirics

Vague probability assessments in national security represent a clear divergence between the theory and practice of high-stakes decision making. As each chapter of this volume argues, evaluating military effectiveness requires judging important tradeoffs under uncertainty. Virtually all national security decisions, at all levels of war, ultimately depend on probability assessments. The consistent practice of leaving those assessments vague deserves scrutiny. This chapter has explored several prominent justifications for that divergence in light of scholarship in security studies and the decision sciences. This review sustains two main conclusions.

First, there is little credible evidence suggesting that expressing probability assessments precisely is unwise or infeasible. Most claims to this effect are not backed by rigorous evidence. Often they are not specified in testable ways. No doubt many national security officials would find it awkward to discuss subjective beliefs using language typically reserved for mathematical reasoning or statistical rigor. But conceding to this aversion has real consequences, adding unnecessary miscommunication to national security debates, and allowing decision makers to evaluate crucial tradeoffs without thoroughly exploring the uncertainty that surrounds them.

This does not mean that proponents of probabilistic precision have an airtight case. Most existing studies of probability assessment involve laboratory settings, undergraduate respondents, and topics unrelated to national security. Only a handful of studies have applied scientific methods to studying probability assessment within the national security domain. The Good Judgment Project is clearly the leader in this field, and its results demonstrate many benefits of quantifying probability assessments. Other scholars can follow that lead. For example, even though the Good Judgment Project has provided unprecedented insight into how foreign policy analysts assess probabilities, there remains a lack of rigorous research investigation how foreign policy officials respond to probabilistic information presented in different forms.

The second main conclusion that this review sustains is that long-standing debates about probability assessment in national security decision making revolve around open, empirical questions. Objections to probabilistic precision are often presented as abstract, epistemological issues, especially the notion that it simply seems inappropriate to discuss subjective judgments using numbers. But analysts always have a valid conceptual basis for quantifying probability assessments if they choose to do so. The question is whether this is actually a good idea – whether quantification biases analysts or decision makers, and if so, whether this bias can be removed. If it is not possible to identify observable empirical negative consequences of stating probability assessments clearly, then it is hard to accept the position that national security officials should continue to provide these assessments in consistently vague ways.

This field of study provides a rich set of opportunities for social scientists. International relations scholars have empirical tools for advancing these debates and a clear theoretical interest in doing so. If it is not a good idea for national security officials to rigorously evaluate probabilities, then it is hard to understand why these assessments deserve such a central place in international relations theory. And if international relations scholars believe that their theoretical frameworks are justified – that any rational evaluation of policy tradeoffs requires carefully assessing probabilities – then it is important to begin closing the gap between theory and practice.