CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

APRIL 2025

Forecasting Nuclear Escalation Risks: Cloudy With a Chance of Fallout

Jamie Kwong, Anna Bartoux, and James M. Acton

Forecasting Nuclear Escalation Risks: Cloudy With a Chance of Fallout

Jamie Kwong, Anna Bartoux, and James M. Acton

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Introduction 1

Exercise Design 4

Escalation Thresholds 7

Key Quantitative Findings 10

Reducing Nuclear Escalation Risks 24

Conclusion and Way Ahead 26

Appendix 1: Workshop Participants 31

Appendix 2: Example Assumptions and Scenario 33

About the Authors 39

Notes 41

Carnegie Endowment for International Peace 43

Introduction

Serious tensions between the United States and China, Russia, and North Korea create the realistic possibility of a high-intensity conventional conflict between two nuclear-armed states. Should such a conflict occur, hundreds of millions of lives could depend on preventing escalation from a conventional conflict to a nuclear war, and, if that failed, from a limited nuclear war to an all-out one.

Forecasting—that is, estimating the probability of specified events' occurring—could contribute to efforts to better understand and address the challenge of managing escalation. In theory, it could be used to estimate the overall risk of nuclear conflict, which should help policymakers decide how much time, money, and political capital they should devote to risk mitigation. Forecasting could also be useful in assessing the potential efficacy of risk-reduction measures, which should help policymakers decide how to invest available resources. Where analysts disagree on the risks of escalation, forecasting may help them to understand why, and thus ultimately narrow those disagreements—which could be especially useful given important debates among analysts about the most likely drivers of escalation.

That said, forecasting nuclear risks—especially risks involving the detonation of a nuclear weapon—poses profound challenges.[1] The employment of nuclear weapons against an adversary is, thankfully, a very rare event. In the eighty years that nuclear weapons have existed, they have only been used in one conflict: World War II. Moreover, the relevance of this case to the contemporary world is highly questionable, not least because Japan's ignorance of the existence of nuclear weapons precluded it from being coerced into surrender by the

possibility of nuclear use by the United States. Since then, there have been only a handful of crises—most famously, the Cuban Missile Crisis—in which the use of nuclear weapons was even remotely plausible.[2] As a result, forecasters are deprived of the data they typically use to calibrate their skills. Given the low probabilities of nuclear-use events, accurate calibration would require a large number of instances of nuclear use and near misses.

The feasibility of forecasting in the face of such uncertainties is far from a purely academic question. In particular, there is evidence that decisionmakers' estimates of the likelihood of nuclear escalation influence their behavior.

According to the journalist Bob Woodward, for example, the U.S. intelligence community assessed in fall 2022 that "if Russian troops were encircled by Ukrainian forces in Kherson, there was a 50 per cent chance [Russian President Vladimir] Putin would order the use of tactical nuclear weapons to avoid such a catastrophic battlefield loss."[3] The fear of a nuclear "Armageddon"—as then U.S. president Joe Biden reportedly described this possibility in October 2022 after being briefed on this intelligence—apparently sparked a diplomatic campaign to warn Moscow of the consequences of nuclear use.[4] In early 2025, then secretary of state Antony Blinken acknowledged such concerns were also partially responsible for the United States' reluctance to supply long-range weapons to Ukraine, stating that "even if the possibility [of nuclear use by Russia] went from 5 to 15 per cent, when it comes to nuclear weapons, nothing is more serious."[5]

Even when policymakers are not provided with probability estimates by intelligence agencies, their thinking may be guided by probabilistic logic. For example, in a deeply researched history of the Cuban Missile Crisis, James Blight and David Welch describe how "hawks" and "doves" focused on the effect of different U.S. courses of action on the probabilities of, respectively, the Soviet Union's removing nuclear missiles from Cuba and escalation to a nuclear war:

> The hawks advocated whichever course of action seemed to have the highest probability of success [in forcing the missiles from Cuba], whether or not it ran a small risk of disaster. . . . In contrast, the doves advocated courses of action which minimized the risk of nuclear catastrophe, whether or not those actions were the most effective.[6]

Given that forecasting—whether or not decisionmakers refer to it as such—influences policy, we conducted a pilot study to systematically investigate its utility in helping to reduce the likelihood and consequences of a nuclear war. This study sought to answer three questions. First, to what extent is there *quantitative* agreement or disagreement among high-skill forecasters and subject matter experts on the probability of nuclear use and subsequent escalation? Quantifying disagreement provides some sense of the uncertainty associated with forecasting. Second, where agreement exists, can forecasting be used to harness expert

judgement and arrive at new *qualitative* insights that ultimately help guide policy? Third, where there is disagreement, can forecasting help to identify its causes and thus enable efforts to resolve it?

Our approach involved three workshops that brought together subject matter experts (of various disciplines) and high-skill forecasters. We asked them to forecast escalation probabilities (as explained below) in specific, hypothetical scenarios of the kind that policymakers might actually face. This approach contrasts with the complementary approach of forecasting events—such as the use of a nuclear weapon or a state's conducting its first nuclear test—that could occur anywhere in the globe over a yearslong time frame.[7]

There is already some evidence of considerable disagreement among expert forecasters about the likelihood that a particular scenario will "go nuclear." In March 2022, for example, eight forecasters with excellent track records were asked to assess the likelihood of dying as a result of the denotation of a Russian nuclear weapon over London in the following month. Their estimates spanned a remarkable seven orders of magnitude (ranging from 2×10^{-2} percent to 5×10^{-9} percent).[8] To our knowledge, however, there has been no attempt to systematically assess whether such uncertainty is endemic to forecasting nuclear use or to compare the forecasts of expert forecasters with subject matter experts about specific scenarios.

We found an extraordinary level of uncertainty in quantitative assessments of escalation risks—uncertainty that stems not primarily from deficiencies in forecasting as a methodology but from its utility in revealing profound disagreements about the phenomenon of escalation. Forecasting can help identify the sources of these disagreements, which include the most likely pathways of escalation; the extent to which the nature of an initial crisis influences the subsequent escalation dynamics; and, to a somewhat lesser extent, the controllability of escalation after first use. Forecasting is also useful for guiding the development of risk-mitigation policies. Our results suggested that efforts to reduce the likelihood of nuclear alerting (preparing nuclear forces so they are ready to launch) and to mitigate escalation pressures after nuclear first use—two generally neglected areas of risk reduction—could play important roles in minimizing the danger of an "all-out" nuclear war.

We conclude by offering suggestions for how forecasters and decisionmakers can increase the utility of forecasting for assessing and mitigating nuclear risks. We urge forecasters to estimate and emphasize the uncertainties in their forecasts. We offer them what we believe to be an honest critique of the strengths and weaknesses of our own methodology so others can improve on it. In parallel, we encourage decisionmakers to make themselves aware of the uncertainty in forecasts of nuclear use.

Exercise Design

We designed a mixed methods exercise that combined a traditional, qualitative discussion around nuclear escalation risks with a quantitative forecasting methodology. To this end, we ran three workshops in 2024. The first was focused on U.S.–North Korean escalation risks (held in Washington, DC), the second on U.S.-Russian risks (London), and the third on U.S.-Chinese risks (Stanford, CA). Each one-day workshop involved a group of twelve to fourteen experts from a range of intellectual disciplines, including expert forecasters, nuclear policy experts, and country specialists, as well as a few psychologists and scientists who had previously applied their expertise to international relations. To the extent possible, we selected subject matter experts with a range of views about the likelihood of escalation and optimal policy responses. Several participants had prior high-level government or military experience (though none were currently serving). In total, thirty-six experts completed the exercise, with one participating in two workshops and one participating in all three (see appendix 1).

Ahead of each workshop, we sent participants four one-page scenarios, each describing the outbreak of a plausible, hypothetical crisis relevant to the dyad of focus. All scenarios took place in the year 2029, and we sent participants a set of assumptions describing, in general terms, relevant countries' military policies, postures, and capabilities in that year. These assumptions could not (and did not try to) completely describe the world of 2029 in all aspects. Indeed, we deliberately refrained from specifying various relevant features, such as which political parties held power in the United States and allied states. At the workshop, we told participants that they should factor such uncertainties into their probability estimates and explained how to do so.[9]

All scenarios involved twenty-five deaths, though the nationalities of the dead and the mix of civilians and military personnel that were killed varied. The scenarios also differed in which party committed the first act of violence and whether the drivers of violence were internal or external, deliberate or unintentional. The scenarios also included relevant U.S. allies and partners—Japan, Taiwan, South Korea, NATO states, and Ukraine, depending on the dyad in question. ("Allies" are countries that the United States is obligated by treaty to defend. The less well-defined category of "partners" comprises states and entities to which the United States is not similarly obligated but might defend anyway and with which Washington cooperates militarily.) All the scenarios and assumptions are available online, and examples of each are included in appendix 2.[10]

During each workshop, the participants were asked to individually estimate the likelihood, for each given scenario, of escalation from each "threshold" to another. There were six such thresholds (see table 1).

Table 1. Crisis and Conflict Thresholds and Their Definitions

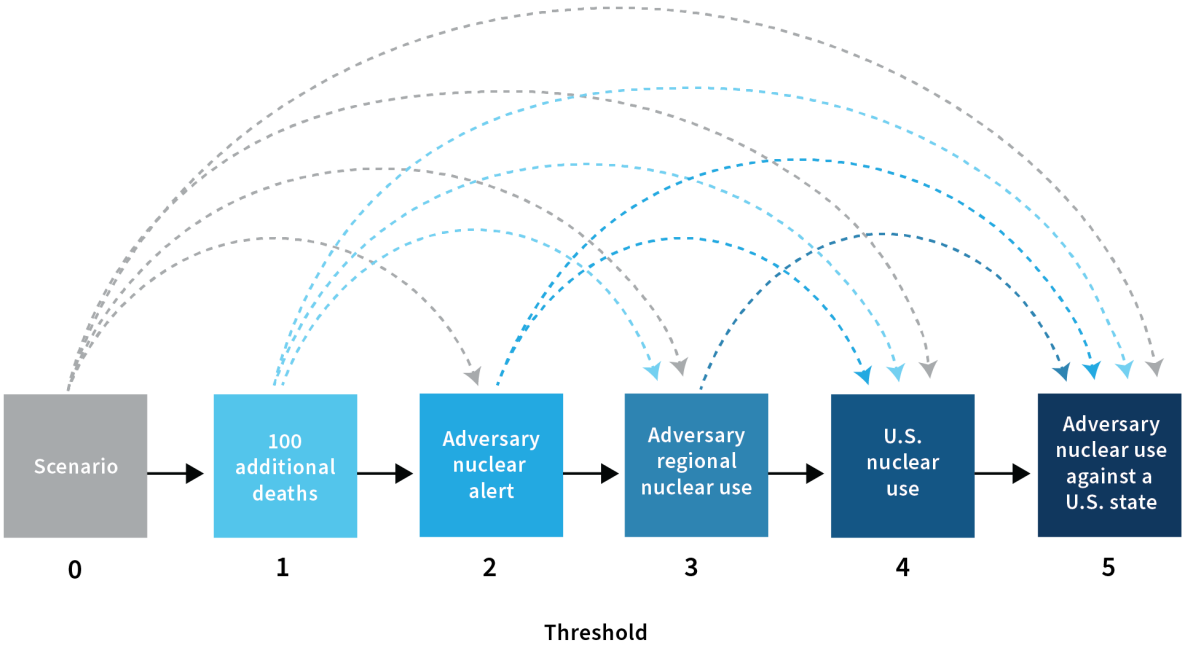
Threshold	Definition
0: Crisis scenario	Described in one-page scenario documents.
1: 100 additional deaths	At least 100 additional deaths (of civilians or military personnel) directly resulting from military action between the adversary and the United States and its allies/partners.
2: Adversary nuclear alert	Relative to the pre-crisis baseline, the adversary removes additional nuclear warheads from centralized storage and/or disperses additional mobile delivery systems armed with nuclear warheads. (The dispersal of nuclear-capable mobile delivery systems that are unarmed or are armed with nonnuclear warheads is NOT considered to be a nuclear alert.)
3: Adversary regional nuclear use	An adversary nuclear warhead (of any yield) detonates (for any reason) anywhere over or inside a regional state, any U.S. territory, or a regional ocean or body of water. The detonation of an adversary nuclear warhead over or inside the adversary's own territory is NOT considered to be adversary regional nuclear use.
4: U.S. nuclear use	A U.S. nuclear warhead (of any yield) detonates (for any reason) anywhere outside of the United States.
5: Adversary nuclear use against a U.S. state	An adversary nuclear warhead (of any yield) detonates anywhere over or inside one of the fifty U.S. states or the District of Columbia.

These definitions do not have a time frame associated with them and, when asked, we defined it as the duration of the crisis or conflict. This approach was intuitive and familiar to nuclear subject matter experts but contrasted with forecasters' expectation of being provided with a fixed time frame.

For each scenario, participants were presented with figure 1. Each arrow represents the conditional probability of escalation from threshold m to threshold n, p(m,n). (A conditional probability is the probability of an event's occurring assuming that some other event has already occurred.[11]) We required participants to estimate the set of probabilities p(m,m+1), that is, the probabilities that the crisis would escalate from each threshold to the one immediately above it (represented schematically by the straight, solid arrows in figure 1).[12] Additionally, participants could choose to estimate any other probabilities (represented schematically by the curved, dashed arrows) that, in their judgment, significantly influenced the overall probability of escalation to the use of nuclear weapons against a U.S. state. We refer to such probabilities—those *not* of the form p(m,m+1)—as "branch" probabilities. The option of including such branch probabilities meant there was no inbuilt assumption that escalation would necessarily occur "linearly" (that is, along each threshold in turn) or without the "skipping" of thresholds.

Figure 1. All the Possible Escalation Steps Between Thresholds

Participants were required to estimate the probabilities associated with the solid arrows. They could choose to provide probability estimates associated with all, some, or none of the dashed "branch" arrows.



We broke consideration of the first scenario at each workshop into a series of segments, each focused on the different ways a given threshold could be crossed. For each segment, we first asked participants to estimate (and write down) the relevant probabilities (in case of crossing threshold 3, for example, those probabilities are p(0,3), p(1,3), and p(2,3), though the only one of these that participants were required to estimate was p(2,3)). Then we facilitated a qualitative discussion of the relevant escalation dynamics, instructing participants not to announce their probability estimates. Our goal was to allow for the pooling of expertise without introducing bias. For the second, third, and fourth scenarios, participants were asked to complete the answer sheet and were then given the opportunity to discuss how the scenarios might differ from one another (again, without announcing their probability estimates). Participants were free to change their estimates of any scenario at any point until submission and were given plenty of time to do so.

After participants submitted their probability estimates, we presented them with various analyses of the results (some of which are summarized below); we allowed them to discuss these results with a particular emphasis on identifying the reasons why they disagreed; and finally, we facilitated a discussion about potential risk-reduction measures.

Escalation Thresholds

Before presenting a summary of our quantitative conclusions, we first summarize the key qualitative issues raised by participants in discussing why thresholds might be crossed and why such a crossing might catalyze further escalation.

This discussion helps to contextualize the quantitative results and could inform future efforts to assess nuclear escalation risks. It also helps illuminate how participants thought about the key drivers of escalation—after all, escalation can usefully be defined as an increase in the intensity or scope of a conflict so it crosses some threshold considered significant by one or more combatants.[13] With one possible exception, all relevant states would view each of the thresholds listed in table 1 as highly significant. The possible exception is the first threshold—100 additional deaths—which is more arbitrary. That threshold was included because it is conceptually useful to demarcate some point at which a crisis escalates into what is, for all intents and purposes, a conventional war.

Threshold 1: 100 Additional Deaths

The workshops highlighted two distinct ways in which participants thought about how the crossing of this threshold could lead to further escalation. At the North Korea and Russia workshops, they tended to focus on who was killed—whether the dead were civilians or military personnel and whether they were citizens of the United States, its allies, or its adversaries—and where the deaths occurred. Some participants argued that the United States and its allies might be particularly sensitive to civilian deaths on their own territory, prompting a more escalatory response. If casualties were primarily among their military personnel, these states might instead undertake what they perceived to be a less escalatory response, such as a high-salience, low-casualty operation (for example, an attack on a symbolic target intended to minimize harm to people).

By contrast, participants at the China workshop generally focused on how the additional deaths occurred. Some argued that it would be much more escalatory, for example, if the United States responded to the initial scenario with one strike that resulted in a high number of Chinese deaths compared to a prolonged, low-lethality exchange that ultimately resulted in the same number of Chinese deaths. These participants tended to posit that the nature of the initial scenarios—whether they reached twenty-five casualties through deliberate or unintended actions—would also inform these responses. Unfortunately, it is difficult to assess the significance of the difference between discussions of U.S.-China contingencies and discussions of scenarios involving Russia or North Korea without conducting additional research to determine whether it persisted when the same scenarios were tackled by different groups of experts.

Threshold 2: Adversary Nuclear Alert

Across the workshops, participants debated numerous reasons an adversary might have for alerting its nuclear forces. Possible motivations identified included: signaling to domestic audiences (to assert authority, for example); coercing an adversary (especially in the case of a conventionally inferior state, to raise the risk of further escalation and thus demonstrate resolve without causing immediate, additional casualties); responding to perceived threats to the survivability of the state's nuclear forces; exacerbating internal political divisions in the United States or its allies; driving a wedge between the United States and its allies; practicing the ability to alert; or catalyzing third parties to involve themselves in the conflict. One major area of disagreement (which is also reflected in the quantitative results) was whether an adversary nuclear alert would be likely to precede or follow the 100 additional deaths threshold.

Threshold 3: Adversary Regional Nuclear Use

Participants at all three workshops generally agreed that the most probable way a conflict would cross the nuclear threshold would be through an adversary nuclear "test" or "demonstration." More specifically, participants at the North Korea workshop typically talked about a North Korean "test," whereas participants in the Russia and China workshops tended to discuss "demonstrations."

Under the deliberately expansive definitions above, any detonation of a nuclear weapon—including a nuclear test in a remote location—is considered to be nuclear use so long as it occurs outside the state's territory.[14] To be sure, there is clearly something unsatisfactory about putting, say, a Russian nuclear test conducted above the Kara Sea in the same category as a nuclear strike on Kyiv. However, in our judgment, trying to distinguish between different kinds of nuclear detonation would be even more methodologically problematic.

Nuclear strategists who regularly use terms like "test," "demonstration," and "attack" may be inclined to disagree. Indeed, those terms certainly have different connotations. A nuclear detonation intended to probe weapon effectiveness would be a test. A detonation intended to signal a state's willingness to use nuclear weapons would be a demonstration if it were conducted in a way that avoided significant physical impact on an adversary. A denotation intended to inflict such an impact on an adversary would be an attack. However, trying to define clearly and unambiguously what kinds of detonations constitute nuclear "use" is fraught. Imagine, for example, a state conducted what it claimed was a "test" by detonating a low-yield nuclear weapon high in the atmosphere over international waters; no test equipment was deployed and the electromagnetic pulse created by the explosion caused moderate damage to electronic systems, but no casualties, in some adversary military vessels that were sailing nearby. Has nuclear "use" occurred? Rather than trying to answer this or similar questions, we felt it was preferable to adopt an expansive but unambiguous definition (though reasonable people can certainly disagree).

In any case, participants disagreed on a number of issues: whether an adversary that had sufficient conventional capabilities to escalate without using nuclear weapons might nonetheless cross the nuclear threshold; the motivations for adversary first use (including the importance of concerns about the survivability of its nuclear forces); the possible targets of such a strike (including differences between military and other targets, and between targets in a U.S. territory in the region and those in the territory of its allies or partners); and the likelihood of inadvertent escalation leading to nuclear use.

Participants also debated how the nature of the original scenario might affect the likelihood that a crisis crossed the nuclear threshold. To give one example, participants in the China workshop generally considered the potential of a conflict to generate threats to the regime—as might occur, for example, if China lost a war over Taiwan—as particularly escalatory (though, notably, they resisted the idea that war aims could expand so that a scenario borne of a different issue might escalate to encompass Taiwan). However, they drew careful distinctions among the natures of various threats to the Chinese regime in assessing the likelihood of Chinese first use. Participants typically argued that a scenario sparked by an internal succession dispute was less escalatory because they expected the United States and its allies would exercise caution amid Beijing's domestic turmoil. Hence, even though the regime was also threatened in this scenario, the potential utility of using nuclear weapons for regime preservation was reduced.

Threshold 4: U.S. Nuclear Use

Participants agreed that the U.S. response to regional nuclear use by an adversary would be highly dependent on circumstances. In all three workshops, for example, participants generally thought it was unlikely that Washington would respond to an adversary nuclear test or demonstration with a nuclear detonation of its own. Outside of this scenario, participants discussed various factors that might inform a U.S. president's decision to use nuclear weapons; ultimately, the president has the sole authority to order such use, creating an inevitable degree of idiosyncrasy in the decisionmaking process. Participants generally agreed that, in assessing a potential nuclear response to adversarial nuclear use, normative pressure not to further erode the nuclear taboo would likely not be a significant consideration. Ensuring the credibility of U.S. security guarantees and declaratory policy might carry more weight, especially in the face of significant pressure from allies. Some participants argued that it would be difficult to envision the president's ordering the use of nuclear weapons except in the case of extreme threats to the United States itself; others asserted it would depend on who is in office.

Some participants argued that the United States might use nuclear weapons first, especially if Washington and its allies were at a conventional disadvantage. Participants could allow for this possibility by assigning significant weight to, for example, p(2,4)—though as discussed further below, they generally did not do so.

Threshold 5: Adversary Nuclear Use Against a U.S. State

Across the workshops, participants debated possible motivations an adversary might have for launching a nuclear attack against a U.S. state, especially given the possibility of a full-scale U.S. response. Many participants agreed that an adversary might be motivated to launch such an attack in response to a nuclear strike on its own territory. They disagreed about whether a conventional attack against an adversary's nuclear assets might also induce it to use nuclear weapons against a U.S. state. Some participants identified crisis instability as a potential motivation (that is, if the adversary came to believe that U.S. strikes might take out most or all of its nuclear forces). Others argued that worst-case assumptions fueled by a lack of information or, relatedly, misperceptions and miscalculations could plausibly motivate an adversary to launch a nuclear attack against a U.S. state. Participants disagreed about the importance of domestic or international political factors.

Some participants argued that adversaries would distinguish between different types of strikes against a U.S. state. For example, in their view, adversaries might believe that the United States would behave differently in response to a high-altitude detonation over Hawaii or Alaska compared to a nuclear attack on, say, Washington, DC.

Key Quantitative Findings

This section presents the key findings from our analysis of the p(m,n) values provided by each participant for each scenario. Participants' estimates for any given p(m,n), as well as the quantities calculated from this data, typically spanned many orders of magnitude. Therefore, unless otherwise stated, we used the median, rather than the mean, as a measure of central tendency and the ratio between the eighty-fifth and fifteenth percentiles to measure the spread. Our notation is summarized in table 2.

Table 2. Notation Summary

p(m,n)	Participant estimate of the probability of escalation from threshold m to threshold n.
PT_n	Probability the conflict reaches and terminates at threshold n (calculated for each participant and scenario from p(m,n) estimates).
$\langle X \rangle_{ps}$	Arithmetic mean of X. The subscript denotes averaging over both participants and scenarios.
$U_{ps}[X]$	Eighty-fifth percentile (upper bound) of X. The subscript denotes that the percentile is taken over both participants and scenarios.
$L_{ps}[X]$	Fifteenth percentile (lower bound) of X. The subscript denotes that the percentile is taken over both participants and scenarios.
R3I	Return on risk reduction investment, defined as the percentage reduction in PT_5 caused by reducing a given p(m,n) by 1 percent.

Finding 1: There were significant disagreements over the likelihood of escalation to a nuclear war and beyond.

The values of p(m,n) supplied by each participant for each scenario can be used to calculate PT_n, the probability that the crisis reaches threshold n and then terminates at that threshold. Thus $PT_3+PT_4+PT_5$ is the probability of a nuclear war, and PT_5 is the probability of an "all-out" nuclear war (or at least what counts as all-out in our simplified, six-threshold escalation model).

Participants disagreed by many orders of magnitude over the probability of both nuclear use and especially an all-out nuclear war, as shown in table 3. To quantify the spread, we calculated the ratio $U_{ps}[P]/L_{ps}[P]$. $U_{ps}[P]$ and $L_{ps}[P]$ are, respectively, the eighty-fifth percentile and fifteenth percentile of the probability, P (either $PT_3+PT_4+PT_5$ or PT_5), taken over all four scenarios and all participants for each dyad. This approach avoids the results being skewed by extreme values. If those extreme values are included, however, the spread becomes even more dramatic. For example, for one of the China scenarios, participants' estimates of PT_5 varied by a scarcely believable *twenty* orders of magnitude. To put this figure into perspective: If the scenario in question were run once per second, one participant's probability estimates suggested that, on average, an observer would witness an all-out nuclear war after just eleven seconds; another's implied that the observer would typically have to wait more than 2,000 times the current age of the universe for the nuclear apocalypse to occur.

Figure 2 offers a visual representation of the spread in data. It shows every participant's estimate of PT_n for each of the scenarios in each of the dyads.

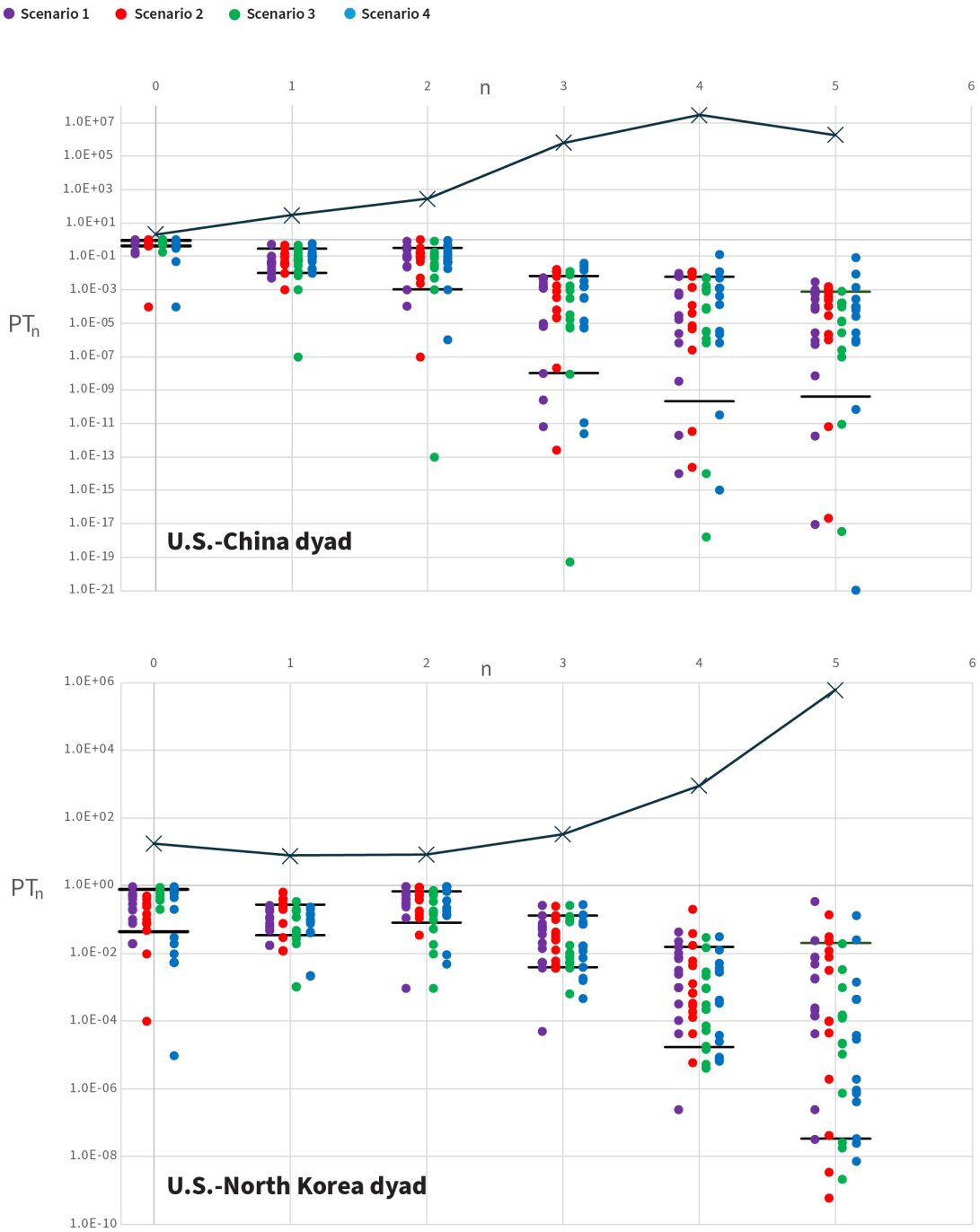
Table 3. Participants' Estimates of the Probability of Nuclear Use and of Nuclear Strikes Against a U.S. State Were Extraordinarily Variable.

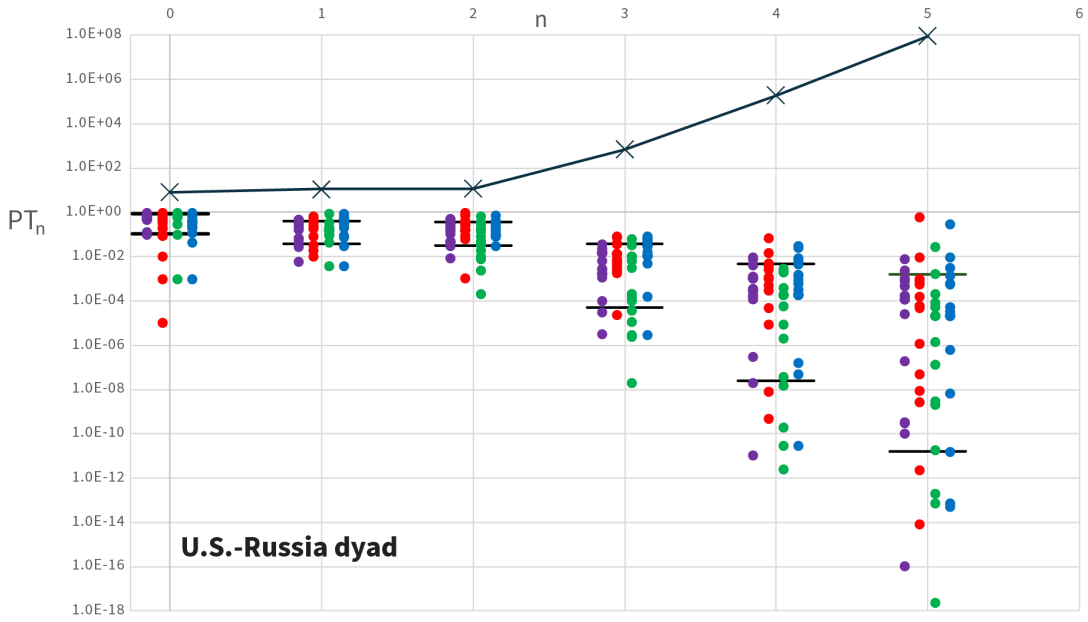
The table shows the fifteenth and eighty-fifth percentiles of the probability estimates of nuclear use ($PT_3 + PT_4 + PT_5$) and nuclear strikes against a U.S. state (PT_5), and the ratios between these percentiles, for each dyad. Because of rounding, dividing the eighty-fifth percentile by the fifteenth percentile may not give the stated ratio exactly.

	$L_{ps}[PT_3 + PT_4 + PT_5]$	$U_{ps}[PT_3 + PT_4 + PT_5)]$	$U_{ps}[PT_3 + PT_4 + PT_5)]/ L_{ps}[PT_3 + PT_4 + PT_5])$	$L_{ps}[PT_5]$	$U_{ps}[PT_5]$	$U_{ps}[PT_5]/ L_{ps}[PT_5]$
China	1×10^{-8}	0.02	1×10^{6}	4×10^{-10}	8×10^{-4}	2×10^{6}
North Korea	7×10^{-3}	0.2	30	3×10^{-8}	0.02	6×10^{5}
Russia	1×10^{-4}	0.05	500	2×10^{-11}	2×10^{-3}	9×10^{7}

Figure 2. The Variation Among Participants' Estimated Escalation Probabilities Generally Increased at Higher Thresholds.

The graphs show participants' estimates of PT_n as a function of n for the U.S.-China dyad (top section), the U.S.-North Korea dyad (middle section), and the U.S.-Russia dyad (bottom section). Each point represents an individual participant's computed estimate of PT_n for a given scenario. The horizontal bars show the eighty-fifth and fifteenth percentiles of these estimates, across all participants and scenarios for each n. The black line shows the ratio between these percentiles.





To be sure, there is probably no meaningful difference between some of the very small estimates of PT_n, especially PT_5. Estimates of, say, $PT_5=10^{-10}$ and $PT_5=10^{-15}$ presumably suggest that the relevant participants think nuclear strikes against a U.S. state are so unbelievably unlikely that they can be discounted entirely.

By contrast, a probability of $PT_5=10^{-5}$ *is* meaningfully different from $PT_5=10^{-10}$. The former would be consistent with a participant's assessment that there was a 10 percent chance of escalation from each threshold to the next along the linear escalation pathway (that is, with all branch probabilities assessed as zero). The latter could indicate a 1 percent chance of escalation at each of these steps. Among skilled geopolitical forecasters, probability estimates for an event that actually has a 10 percent chance of occurring are generally meaningfully different from probability estimates for an event that actually has a 1 percent chance of occurring.[15] The forecasts of subject matter experts are unlikely to be so well calibrated; even so, among those individuals, it still seems highly likely that an estimate of $PT_5=10^{-5}$ is meaningfully different from, say, $PT_5=10^{-15}$.

Before running the exercise, we had hoped to be able to compare the level of danger between different kinds of scenarios. For example, is escalation in scenarios involving North Korea more likely than in those involving China or Russia? Are scenarios catalyzed by internal events, such as leadership transitions, more dangerous than those driven by external factors? Because of the extraordinary degree of uncertainty in PT_n, however, no such comparisons were anywhere close to being statistically significant.

Finding 2: Participants were surprisingly confident about how escalation might occur—but disagreed about which escalation pathways were the most likely.

During workshop discussions, most participants emphasized (in, we believe, entirely good faith) the uncertainty surrounding escalation. In particular, even without prompting, many acknowledged real difficulties in assessing how escalation might unfold—let alone in assigning probabilities—due to the lack of clarity in the exercise around states' intentions, motivations, and goals, even while acknowledging that this lack of clarity is a realistic part of assessing real-world nuclear escalation risks. Analysis of their probability estimates, however, reveals a different story: a surprising degree of confidence in how escalation might unfold.

An "escalation pathway" refers to the ordered crossing of some set of thresholds. For example, if escalation culminates with adversary nuclear use against a U.S. state, then our model allows for sixteen different pathways. The initial scenario, threshold 0, could (at least in theory) escalate straight to nuclear strikes against a U.S. state without any intermediate steps (we denote this possibility as 0-5). Alternatively, it could escalate to 100 additional deaths, before the adversary launched nuclear attacks against a U.S. state (0-1-5). Or the initial crisis could instead be followed by an adversary nuclear alert and then by strikes against a U.S. state (0-2-5), and so on, all the way through to the linear pathway (0-1-2-3-4-5).[16]

For a given scenario, we use the term "principal pathway" to denote a participant's assessment of the escalation pathway most likely to be followed *in a conflict that culminated with the adversary's ultimately launching nuclear strikes against a U.S. state*. Or, to put it another way, of all those sixteen escalation pathways that ultimately end at threshold 5, which did a given participant assess to be the most likely?

The concept of the principal pathway is useful because, in general, participants' results implied that all-out escalation is likely to occur along the principal pathway; by contrast, participants' typical qualitative comments about the uncertainty associated with escalation suggested that escalation along the principal pathway should be only slightly more likely than along other pathways.[17]

The size of this discrepancy can be quantified. On average, participants assessed that, in the event of an all-out nuclear war, there was a probability of about 0.70–0.75 that it would result from escalation along the principal pathway (see table 4, which, because all estimates were of the same order of magnitude, presents their arithmetic mean, rather than their median).[18] To put this finding in context, with sixteen available pathways, this probability could be as low as 0.0625 (which would occur if all the pathways were equally likely). The figures in table 4 are also striking because of their consistency between dyads, in contrast to much else reported in this paper.

Table 4. Participants Believed That the Importance of the Principal Pathway Varied Little Between Dyads.

The table shows the mean calculated probability that escalation would occur along the principal pathway for each dyad. The uncertainty estimate is the standard error.

	$\langle P(\text{escalation occurring along the principal pathway}) \rangle_{ps}$
China	0.69 ± 0.04
North Korea	0.76 ± 0.03
Russia	0.74 ± 0.03

To be sure, it is unclear which is correct: the quantitative finding that escalation to an all-out nuclear war is likely to proceed along the principal pathway or the intuitive expectation that a number of pathways should have broadly similar probabilities. Rather, our main aim here is to draw attention to the discrepancy and encourage efforts to probe it further. That said, there is one reason to at least question the quantitative finding. Although data from individual participants implied confidence in how escalation might proceed, there was little agreement about what the principal pathway was. For each of the scenarios, participants identified at least five, and as many as eight, different principal pathways.

To give a particularly extreme example, in one of the North Korea scenarios, a participant identified the principal pathway to be 0-5 and assessed that, in the event of nuclear strikes on a U.S. state, there would be a 0.99999 chance that escalation would proceed along that pathway. Conversely, for the same scenario, another participant identified the principal pathway as 0-1-2-3-4-5 and assessed that all-out escalation was certain to proceed along it. Clearly, it is impossible for both these participants to be correct.

Finding 3: Participants were generally—though not universally—optimistic about managing escalation after first use.

Analysts can have a variety of views about the prospects for managing escalation after the first use of nuclear weapons. At one extreme, "optimists" believe that there would be a reasonable prospect of controlling escalation. In other words, escalation after first use would be sufficiently unlikely that $PT_3 > PT_4 > PT_5$. At the other extreme, "pessimists" believe that, if nuclear weapons were used in even a limited way, the ultimate result would likely be all-out escalation, so $PT_3 < PT_4 < PT_5$. There are other possibilities too, of course. In fact, six beliefs can be imagined based on pairwise comparisons of PT_3, PT_4, and PT_5 (see figure 3). We

assign each belief a letter, with type O being the most optimistic view, type P the most pessimistic, and the other possibilities being labelled arbitrarily types A through D. For example, type B indicates the belief that there would be a reasonable prospect of preventing a nuclear war from escalating from adversary regional nuclear use to U.S. nuclear use, but that if the United States used nuclear weapons, the result would likely be all-out escalation.

Across all dyads, participants, and scenarios, all six different beliefs about nuclear escalation were observed, but they were not equally common (see table 5). In about two-thirds of cases, participants exhibited type O beliefs. The next most common were type A (19 percent) and type C (10 percent). The former reflects a situation in which the adversary is more willing to engage in nuclear brinkmanship than the United States. The latter reflects a situation in which the United States is more willing to engage in brinkmanship than the adversary. The overall consistency between the workshops concerning which pathways were the most common is notable—and it would be interesting to see if it were retained across additional experiments.

Figure 3. Six Possible Beliefs About Escalation Within a Nuclear War.

The figure shows schematically PT_n as a function of n for each belief type, as well as their definitions.

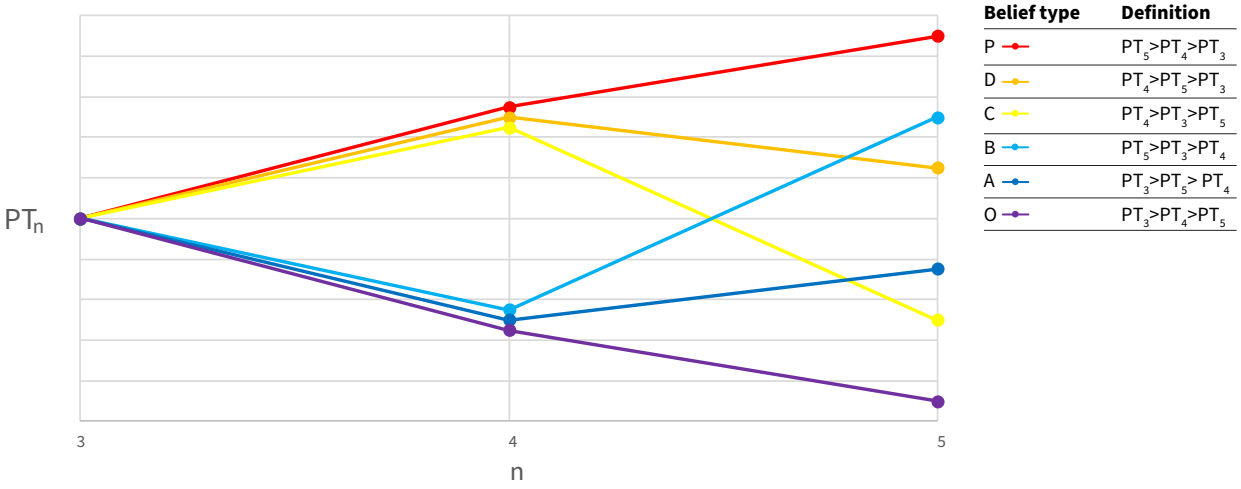


Table 5. Most, But Not All, Participants Were Optimistic About the Prospects for Escalation Management After Nuclear First Use.

The table shows the frequency of belief types about the controllability of escalation after nuclear first use among participants for all scenarios in each dyad. The belief types are defined in figure 3.

Belief type	China (%)	North Korea (%)	Russia (%)	Total (%)
O	50	65	73	63
A	21	19	18	19
C	19	8	4	10
P	8	0	0	3
D	2	4	2	3
B	0	4	4	3

These findings are at least slightly surprising given the qualitative discussions around adversary capabilities and risk tolerance—with the caveats that these discussions focused more on assessing the prospect of first use than on managing escalation *after* first use and that there are relatively modest differences between the results from the workshops.

Participants consistently noted the relevance of the conventional military balance in considering the potential for nuclear use. In the China workshop, some participants argued that because the conventional balance is likely to tilt in Beijing's favor by 2029, it would likely not need to use nuclear weapons to achieve its aims. By contrast, participants assessed that North Korea and Russia may rely more on their nuclear weapons as the conventional balance in 2029 is likely to tilt in favor of the U.S.–South Korea alliance and NATO, respectively.

Moreover, participants at the China workshop generally implied that Beijing would be fairly risk averse (even if this assumption was rarely stated explicitly). They believed that China would likely avoid entering into crises serious enough to risk nuclear escalation in the first place. Comparatively, participants at the North Korea and Russia workshops seemed to assume that both states would have a greater tolerance for risk.

Based on these assessments, one might expect the China workshop results to have had the highest percentage of optimists about escalation management rather than the lowest. The prospect of a military rebalance in the Indo-Pacific that tilts in China's favor, however, does seem to align with a U.S. willingness to engage in nuclear brinkmanship, which is consistent with the relatively high percentage of type C beliefs in this case.

Finding 4: Participants disagreed strongly on the extent to which escalation dynamics after first use were influenced by the nature of the initial crisis.

Participants' probability estimates can be used to infer their beliefs about the extent to which the initial crisis affects escalation dynamics after the first use of nuclear weapons. In theory, a variety of such beliefs are possible. At one extreme, an analyst might argue that such dynamics would depend sensitively on the crisis that had precipitated the nuclear war; states would continue to pursue their political goals and their willingness to engage in brinkmanship would depend significantly (though perhaps not exclusively) on their stake. At the other extreme, an analyst might argue that once nuclear weapons had been used, the crisis would be transformed; with national survival—and indeed the survival of civilization itself—at stake, the original crisis would be rendered essentially irrelevant. While we did not ask participants directly about their beliefs, their probability estimates provided an indirect means of assessing those beliefs. A participant who predicted similar escalation dynamics after nuclear use across all four scenarios presumably believes that the initial scenario is largely irrelevant. Conversely, if the dynamics varied with the scenario, the participant presumably believes that the original scenario is important.

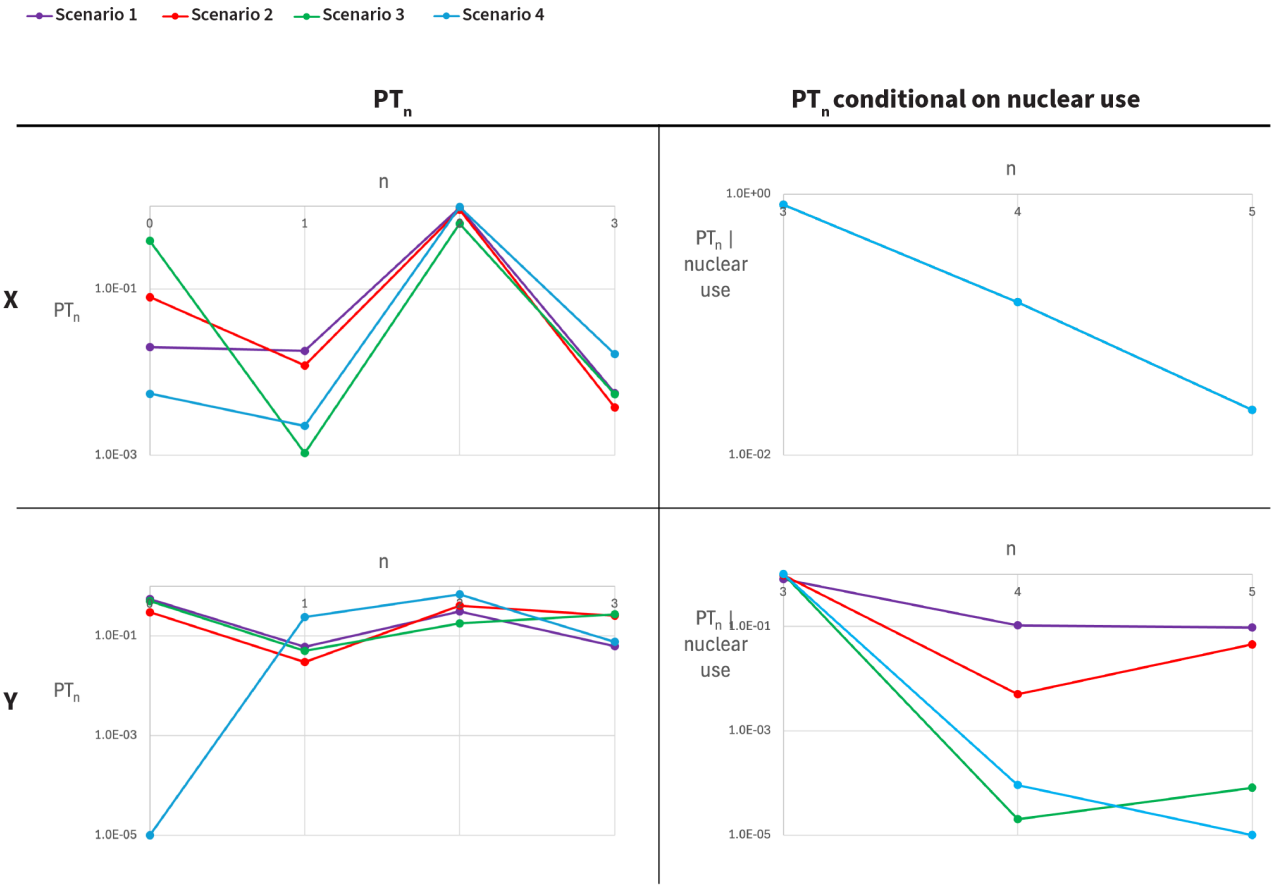
Participants had a variety of opinions about the extent to which the nature of the initial crisis affects escalation dynamics after first use. One indication is that about half of all participants (eighteen out of thirty-nine) had the same belief about the controllability of nuclear escalation in all four scenarios. Of the remaining twenty-one participants, data from nineteen revealed two different beliefs across the four scenarios, and data from the other two revealed three beliefs.

Calculations of PT_n and, separately, PT_n *conditional on nuclear use* support a similar conclusion. (The latter quantity represents the probability that a crisis terminates at threshold n, assuming that it ultimately reaches threshold 3, 4, or 5.)

To give a concrete example, figure 4 shows such data for two participants (labeled X and Y) from the North Korea workshop. Participant X expected the four North Korea scenarios to unfold quite differently prior to nuclear use (as can be seen by the divergences of the lines in the upper left section of the figure). However, in the event of nuclear use (that is, conditional on threshold 3, 4, or 5 being reached), this expert judged the probabilities of further escalation to be identical (as can be seen by the fact that, in the upper right section, the four lines lie exactly on top of another, so only one is visible). In other words, participant X believes that the nature of the initial crisis is important to determining escalation dynamics before first use but becomes entirely irrelevant afterwards. Participant Y has an almost diametrically opposed view. This expert judged that three of four scenarios would unfold almost identically prior to first use (bottom left section), but that important differences between the scenarios would emerge after first use (bottom right section). Indeed, participant Y exhibited three different kinds of beliefs about the controllability of nuclear war across the four scenarios.

Figure 4. One Participant Believed That Escalation Dynamics Before Nuclear First Use Would Depend on the Initial Scenario but Those Afterwards Would Not; Another Participant Had a Diametrically Opposite View.

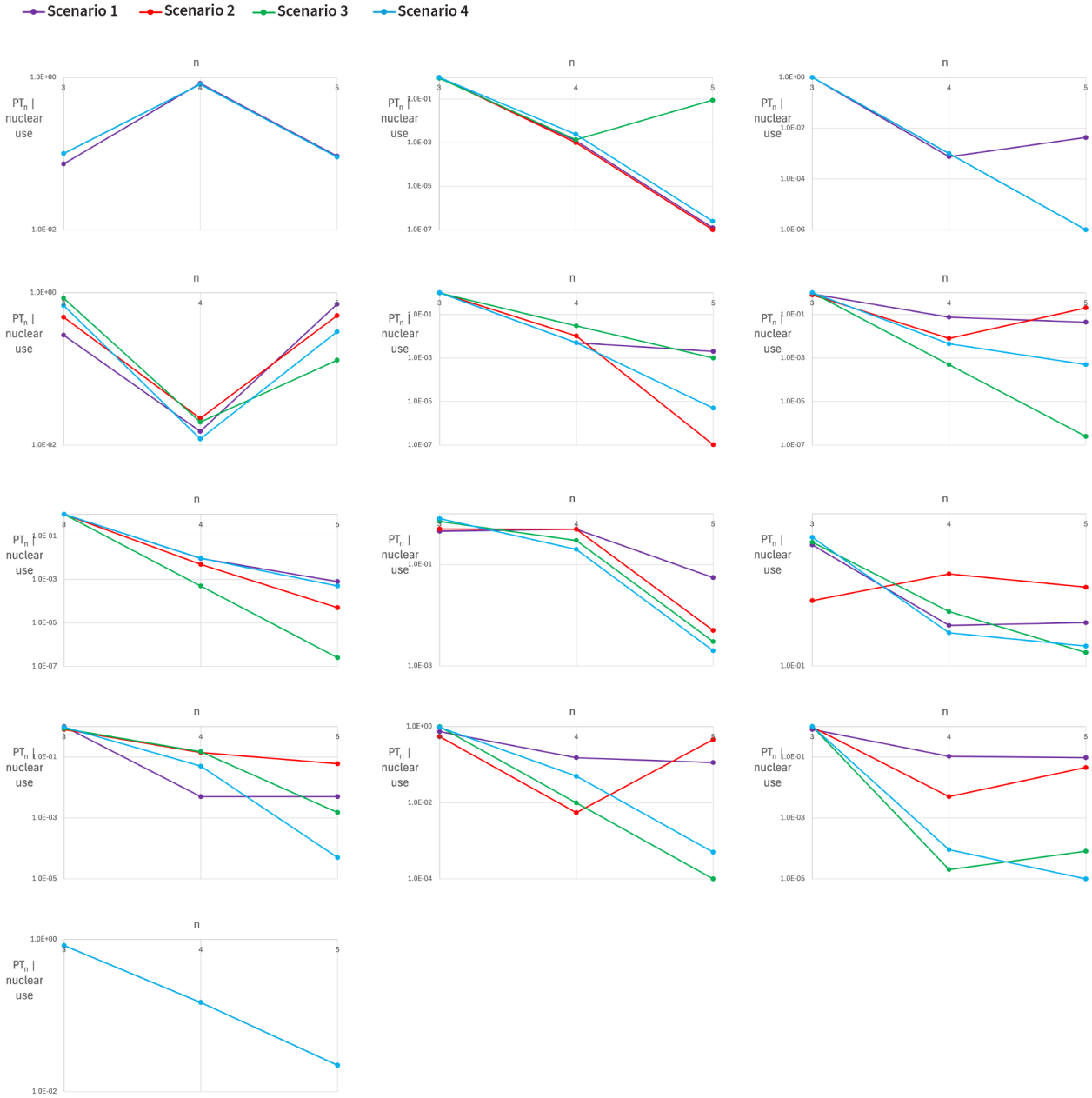
The table shows the calculated estimates for two participants, X (upper section) and Y (lower section), of PT_n (left section for n=0, 1, 2, and 3) and PT_n conditional on nuclear use (right section for n=3, 4, and 5) for each North Korea scenario. In the upper right section, the graphs for each scenario are coincident.



In fact, across all participants in the North Korea workshop, there was a wide variation in the extent to which escalation dynamics after first use depended on the initial crisis, as shown in figure 5. A similar, if somewhat less dramatic, dependence is visible in the results from the China and Russia workshops, but for brevity, only the North Korea results are provided here.

Figure 5. Participants Had a Wide Variety of Beliefs About the Extent to Which the Initial Scenario Would Affect Escalation Dynamics After Nuclear First Use.

The graphs show the calculated estimates for all participants of PT_n conditional on nuclear use as a function of n (for n=3, 4, and 5) for each North Korea scenario.



Finding 5: There was no relationship between area of expertise and views about nuclear war.

Previous forecasting studies have found that subject matter experts tend to be more pessimistic than expert forecasters about the likelihood of events they study.[19] In our work, there was no relationship at all between expertise and estimates of the likelihood of escalation (which is, perhaps, not surprising given the enormous variations in such estimates).

Qualitatively, expertise did, however, affect some of the discussions around exercise design. For example, nuclear experts generally expected a conflict to play out quickly, and some stated that the inclusion of a timeframe of say, one year, in our threshold definitions would not have significantly altered their probability estimates. By contrast, forecasters mostly expected events to play out more slowly and indicated that a timeframe of one year would have led them to lower their probability estimates.

Finding 6: In theory, the most effective way to reduce the likelihood of all-out escalation is to reduce the probability of escalation from one threshold to the one immediately above it, especially at steps higher up the escalation ladder. In practice, identifying means to reduce those probabilities is difficult. Preventing nuclear alerting early in a crisis may present a more feasible and somewhat overlooked opportunity for risk reduction.

Return on risk reduction investment (R3I) is defined as the percentage reduction in PT_5 caused by a 1 percent reduction of any given p(m,n). It can help guide policy by identifying which steps in the escalation ladder are most consequential. Moreover, even if analysts radically disagree over the likelihood of escalation, they may still agree on which p(m,n) offers the best R3I. Table 6 shows the rankings of the median R3Is (across all participants and scenarios) at each workshop as well as an average ranking over the three workshops. Figure 6 shows the absolute values of the median R3I rankings (again, across all participants and scenarios) for the China workshop (similar figures for the Russia and North Korea workshops are omitted for brevity).

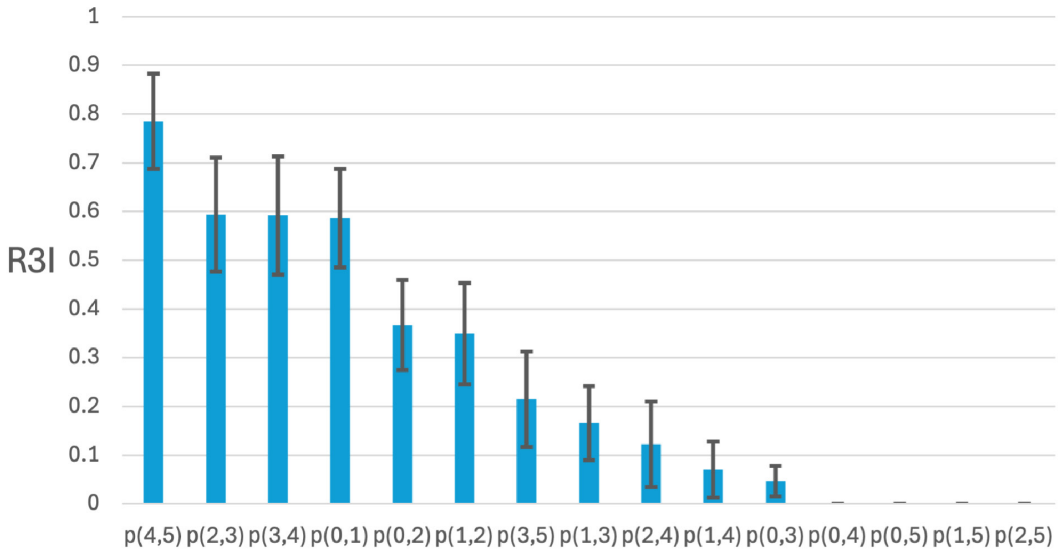
Table 6. The R3I Rankings Were Generally Consistent Across the Dyads.

The table shows rankings of the median R3I for each p(m,n) for each dyad, along with the average ranking over all three dyads

	China	Russia	North Korea	Average ranking
p(4,5)	1	1	1	1
p(3,4)	3	2	2	2.3
p(2,3)	2	3	4	3
p(0,1)	4	4	3	3.7
p(0,2)	5	6	5	5.3
p(1,2)	6	5	6	5.7
p(3,5)	7	8	7	7.3
p(0,3)	11	10	8	9.7
p(1,3)	8	11	11	10
p(2,4)	9	7	14	10
p(1,4)	10	9	15	11.3
p(0,5)	13	13	10	12
p(0,4)	12	12	13	12.3
p(1,5)	14	14	9	12.3
p(2,5)	15	15	12	14

Figure 6. Steps Between Adjacent Thresholds Near the Top of the Escalation Ladder Had the Highest R3Is.

Median R3I for each p(m,n), across all participants and scenarios, for the U.S.-China dyad. Uncertainty estimates span the fifteenth to eighty-fifth percentiles.



One interesting feature of the median R3I rankings is the relative consistency between workshops, which is somewhat unexpected given the disagreement on other issues. To be fair, that consistency stems, in part, from defining R3I in terms of reducing PT_5. As a result, steps higher up the ladder—such as escalation from threshold 3 to 4, or from 4 to 5—tend to offer greater R3Is because they would be taken if escalation to an all-out nuclear war proceeded in a linear way (as many participants thought it would).

That said, this is not the whole story. It was not a priori obvious that there would be a low R3I to efforts to reduce the danger of a conflict's escalating from a relatively low level (threshold 0, 1, or 2) straight to nuclear strikes on a U.S. state. To be sure, the relevant probabilities—p(0,5), p(1,5), and p(2,5)—are inevitably very small in absolute terms (surely much smaller than those between sequential escalation thresholds). However, potentially counteracting the small size of these probabilities is that escalation pathways that skip multiple thresholds involve fewer steps (in mathematical terms, the likelihood of a multiple-step escalation pathway is calculated by multiplying the probability associated with each individual step). As a result, it is possible for the highest R3Is to come from reducing the probabilities p(0,5), p(1,5), and p(2,5), as indeed occurred with a small number of participants.

On average, however, the highest R3Is are associated with p(4,5), p(3,4), and p(2,3). In other words, in theory, the most effective way to prevent nuclear strikes against the United States would be to reduce the likelihood of escalation from U.S. nuclear use to such strikes; from adversary nuclear use in the region to U.S. nuclear use; and from an adversary nuclear alert to adversary regional nuclear use. However, this is often easier said than done. As discussed below, very few of the policy recommendations suggested by participants related to curtailing escalation risks after first use; in line with the post–Cold War nuclear discourse, more recommendations related to preventing first use itself.

The data, however, identify one promising target for risk reduction: reducing the likelihood that an adversary's nuclear forces are alerted early in a crisis. Many participants believe that such an adversary alert would occur before the 100 deaths threshold was reached. This expectation is reflected by the relatively high R3I of p(0,2).

The value of reducing p(0,2) can be seen by calculating the estimated probability that, conditional on an adversary's alerting its nuclear forces, the total number of deaths is fewer than 100. Results were surprisingly consistent between the workshops (see table 7, which, because of the relative consistency of probability estimates within a dyad, gives the arithmetic mean of this probability rather than the median). Participants' p(m,n) estimates implied that, if an adversary alerts its nuclear forces, there is a probability of about 0.4-0.45 that fewer than 100 deaths have occurred.

Table 7. Participants Believed That Nuclear Alerting Might Well Precede 100 Additional Deaths.

The table shows the mean probability of fewer than 100 additional deaths given a U.S. adversary alerts its nuclear forces. The uncertainty estimate is the standard error.

	<P (fewer than 100 additional deaths given adversary nuclear alert)>$_{ps}$
China	0.45 ± 0.05
North Korea	0.44 ± 0.05
Russia	0.42 ± 0.04

Aiming to reduce p(0,2) is an attractive prospect because, compared to steps later in the escalation ladder, the alerting of nuclear forces is *relatively* well understood, not least because there have been various historical instances of alerting. Analysts may therefore consider giving particular attention to nuclear alerting in crafting future risk-reduction efforts.

Reducing Nuclear Escalation Risks

Toward the end of our first workshop, which was focused on North Korea, we asked participants to suggest one risk-reduction measure that would address the escalation risks they had spent the day analyzing. It was notable how few of those measures corresponded to the escalation steps with the highest R3Is. At subsequent meetings, we told each participant which escalation step had the highest R3I, according to their individual probability estimates, and asked them to identify one practical way to reduce the probability of its occurring. They often struggled to do so, especially when the relevant step was high up the escalation ladder. Indeed, recommendations were generally divorced from probability estimates; participants likely offered the same recommendations that they would have suggested prior to the exercise—a notable finding in and of itself and another example of the disjuncture between the qualitative and quantitative.

Across the workshops, recommendations focused on four general themes (all of which are summarized in table 8):

- **Posture and capabilities:** Participants proposed various changes to U.S. nuclear policy (which could be implemented unilaterally). Proposed changes generally focused on declaratory policy (such as adopting a sole purpose doctrine or acknowledging mutual vulnerability with China) and capabilities (such as investing in point and area missile defenses and improving nonnuclear capabilities).

- **Influencing adversary perceptions**: Some recommendations focused on shaping adversary perceptions in a crisis to demonstrate political resolve (such as signaling through regional deployments or capability investments) or alliance cohesion (such as through coordinated strategic messaging) and thus enhance the United States' deterrence credibility. By contrast, others focused on adversary reassurance (such as efforts to reassure North Korea that if it does not use nuclear weapons, the United States and South Korea will not try to end its regime).

- **Cooperative risk-reduction measures:** Crisis prevention and communication tools—notably hotlines—featured prominently at all workshops. Participants recognized that communication mechanisms are highly imperfect: adversaries may refuse to use them; in a conflict, they might be disrupted by conventional or nuclear operations; and adversaries may not trust whatever information is conveyed. Even so, participants generally agreed that their benefits outweigh the risks.

- **Enhanced understanding of escalation management:** Participants recommended various analytical efforts officials and nongovernmental experts could take—ranging from improved scenario planning to better analysis of adversaries—to equip the United States with the knowledge and insight needed to manage escalation more effectively.

Table 8. Policy Recommendations: Themes and Examples

Posture and Capabilities	Influencing Adversary Perceptions	Cooperative Risk-Reduction Measures	Enhanced Understanding of Escalation Management
Changes to declaratory policy (for example, sole purpose) Changes to sole authority (the principle that a U.S. president can authorize nuclear use unilaterally) Capability investments (for example, point and area missile defenses) Capabilities to improve U.S. resilience across domains (for example, cyberspace, space) Enhanced allied conventional capabilities	Increased deterrence credibility through political means (for example, signaling though regional deployments) and military capability investments Demonstrate alliance cohesion (for example, coordinated strategic messaging) Enhanced external pressure through a third party Reassurance that certain targets are off-limits if an adversary refrains from particular action	Crisis communication tools (especially hotlines) Launch notifications Failsafe reviews Track 1, 1.5, and 2 dialogues	Scenario and response planning Improved adversary analysis Improved wargaming Limited nuclear use plans

Underlying this list was, inevitably, considerable disagreement. Some measures, if not exactly mutually exclusive, are certainly in tension. For example, there was broad agreement that efforts to deter and assure adversaries generally cut against one another. In other cases, there was disagreement about whether there were, in fact, trade-offs between different measures and, if so, how severe they were. For instance, participants debated the interplay between diplomacy focused on crisis prevention and efforts to enhance (conventional and nuclear) capabilities that might be useful for escalation management.

The policies advocated by participants generally depended on what they saw as the key escalation drivers. For instance, participants at the Russia workshop who argued that escalation could happen because the United States and NATO do not properly understand Russian intentions tended to focus on improving communications and better understanding how Russian elites think about escalation and nuclear use. By contrast, those who argued that the United States and NATO are not doing enough to deter Russia tended to focus on increasing deterrence credibility.

These various debates are not novel; they will be intimately familiar to nuclear policy experts. Indeed, it was somewhat disappointing that the forecasting framework did not do much to advance the debate over risk mitigation—an issue we consider in the next section.

Conclusion and Way Ahead

After reading about the challenges encountered in our attempt to forecast nuclear use events, an understandable reaction would be to throw in the towel—to argue that disagreement among experts results in "answers" that are so uncertain that they are useless and, indeed, that the whole idea of forecasting nuclear use events is so flawed that it should be abandoned.

This response, however, would be a mistake. The problem is not that forecasting is an inappropriate tool to understand escalation. Rather, the uncertainty stems from unanswered (and perhaps unanswerable) questions about escalation as a phenomenon. In other words, uncertainty is not a methodological artifact of forecasting. It comes from profound disagreements between informed, experienced, and skilled experts about how escalation would actually unfold in a crisis or conflict, especially one that was close to or beyond the nuclear threshold. Forecasting offers a way to understand these disagreements and to derive potentially useful outputs despite the uncertainty.

The first benefit of forecasting is simply to highlight the extent of the uncertainty surrounding escalation. To be sure, most experts would acknowledge this uncertainty—at least in the abstract. Yet, the magnitude of the uncertainty captured in our study is shocking. Dismissing forecasting may make life easier for both policymaker and analyst by allowing them to ignore how poorly understood the phenomenon of escalation really is. Any resulting increase in confidence, however, would be ill-founded.

Even if forecasting were useful for nothing more than underscoring uncertainty, it would have real value. However, we believe, albeit very tentatively, that it can provide useful qualitative insights. For example, our results highlight two areas that have probably received insufficient attention as part of risk-reduction efforts: nuclear alerting and escalation after nuclear first use.

Forecasting offers two other benefits. First, as many others have noted, it can help analysts to understand why they disagree. Specifically, our exercise revealed important disagreements about the most likely pathways of escalation, the effect of the nature of the initial crisis on subsequent escalation dynamics, and, to a somewhat lesser extent, the controllability of escalation after first use. It was less useful in understanding why participants disagreed so strongly about the overall likelihood of escalation. In theory, this deficiency could be addressed by decomposing any given step in our escalation ladder into a series of pathways and estimating their probabilities. In practice, while potentially valuable, this process would be difficult and time-consuming—too time-consuming, certainly, for a one-day workshop that aimed to explore an escalation ladder that began with a crisis and ended with nuclear strikes against a U.S. state.

Second, forecasting can raise "red flags" by identifying differences between qualitative expectations and quantitative assessments. For example, it highlighted the tension between participants' statements that it was difficult to predict how a crisis might escalate and their assessment—inferred from probability estimates—that if escalation occurred, it was likely to proceed along the principal pathway. In the event of such a difference, there is no a priori way of knowing whether the quantitative assessments or qualitative expectations are more reliable; instead, the point is simply that further research is needed.

Observing these benefits (or at least potential benefits) is not to suggest that all is rosy in the forecasting garden. Far from it. Forecasting nuclear use is profoundly difficult. To increase its utility, we offer suggestions for forecasters and forecasting (focusing on methodology) and suggestions for policymakers (focusing on interpretation).

Suggestions for Forecasters and Forecasting

Most importantly, we urge forecasters to emphasize the inevitable uncertainty in their estimates of nuclear use probabilities. Ignoring or downplaying these uncertainties risks giving policymakers false confidence. Quoting nuclear-use probabilities to two or three significant figures, for example, gives the impression that escalation is a much better understood phenomenon that it actually is. To rectify this problem, forecasters should do more than simply cite their estimates to one significant figure (or perhaps just to the nearest order of magnitude). They should openly and prominently acknowledge the uncertainty in their estimates and, where possible, provide quantitative estimates of it.

Even as much of this uncertainty stems from the underlying phenomenon, methodological improvements in forecasting studies of nuclear use could help to reduce it. Indeed, we encourage others to assess our methodology and test improvements. To help, we offer our own critique and suggestions for further work.

Inevitably, we had to make various hard calls on methodology, primarily because we faced significant trade-offs between complexity and methodological fidelity. The primary challenge is that probabilities are not intuitive to all subject matter experts, many of whom have little or no quantitative training. Not all workshop participants understood the concept of a conditional probability at the start of each workshop; while we explained and practiced this concept prior to any forecasting, some still struggled with the exercise.

Anticipating these challenges, we decided against requiring participants to estimate branch probabilities, which may have biased them in favor of linear escalation. We also refrained from asking participants to assign uncertainty estimates to their probabilities. In our judgment, requiring participants to assign fifteen probabilities (one for each arrow in figure 1) for each scenario, let alone adding uncertainty estimates to those probabilities, would have added significantly more difficulty to an already complex task, making the whole exercise potentially unmanageable. We continue to believe these calls were correct.

By contrast, we regret basing each workshop on a different dyad. With the benefit of hindsight, we would have focused our pilot study on just one dyad and run the same four scenarios at all three workshops so their results were directly comparable. Future work could correct this problem.

We made a series of other methodological choices about which reasonable people can differ. For example, we instructed participants to estimate escalation probabilities over the duration of the crisis rather than in a fixed timeframe. We defined nuclear use broadly so that it included any nuclear detonation, including a test or demonstration, beyond a state's territory. It would be interesting and worthwhile to determine whether different choices would change the results significantly.

We also asked participants not to share their probability estimates as they discussed how escalation might occur because we did not want them to anchor onto one another's estimates. There would be great value, however, to testing an alternative approach in which participants worked as small teams or else revealed and debated their probability estimates. In doing so, at least two challenges would need to be overcome. First, and most practically, the alternative approach would likely be more time-consuming. Second, there is a danger that participants would change their probabilities out of social pressure to conform or if they lacked confidence in making probabilistic assessments, and not because they had been genuinely persuaded by others. In that case, some or all the reduction in uncertainty would be illusory.

An ambitious goal could be to try to combine forecasting and simulations. In wargames and tabletop exercises, which are used widely within the national security community, teams representing the leadership of two antagonists are presented with a scenario and then take actions in response to one another. As part of such an event, the participants could (individually or collectively) be asked to estimate probabilities of escalation, including nuclear use. It could be useful to analyze how these probabilities change over the course of a simulation and to compare the two teams' estimates.

Finally, future work could facilitate more discussion among participants about important differences between their qualitative expectations and quantitative assessments. Such discussions were not prominent in our workshops, in part because we only identified some of these differences when analyzing data after the workshops. That said, even if we had identified all these differences in real time, it would have been difficult to facilitate an in-depth discussion given time constraints.

Suggestions for Decisionmakers

In a crisis or conflict, intelligence agencies may present decisionmakers with estimates (perhaps conditional) of the likelihood of nuclear use by an adversary. Decisionmakers may also develop their own such estimates, explicitly or implicitly. Such estimates are potentially useful as one input (of many) to policymaking, especially if decisionmakers are willing to look beyond the topline to consider uncertainty and its causes.

Once again, this is probably easier said than done. Few decisionmakers are comfortable with probabilistic thinking.[20] Moreover, in a crisis or conflict, time is likely to be of the essence; realistically, no decisionmaker can spend hours debating a complex probability tree. On top of that, key advisers may want to shield their principals from uncertainty to simplify decisionmaking. Yet, for their part, within all the real and inevitable constraints, decisionmakers should want to be fully informed; false confidence in the likelihood of nuclear escalation is not a firm foundation for policy. To this end, we offer two simple recommendations.

First, decisionmakers should require all quantitative forecasts developed by intelligence agencies to contain uncertainty estimates. (The confidence assessments that are already included in U.S. intelligence estimates are a qualitative way of expressing the probability of a given outcome; the proposal here is for a clear statement of the uncertainty in such estimates.) Perhaps the simplest approach is to provide decisionmakers with the range of individual estimates that fed into a consensus forecast (though enhanced methods could undoubtedly be developed). Ideally, if decisionmakers developed their own forecasts, they would also think about uncertainty, though from our experience, most people without extensive mathematical training struggle to estimate the uncertainty in probabilities. Nonetheless, decisionmakers could still ask themselves, "if my estimate is wrong, what are the most likely reasons why?" as a more qualitative approach.

Second, regardless of whether there is significant uncertainty in probabilities estimates—though, in the case of nuclear use, there normally would be—decisionmakers could ask advisers what kind of escalation pathway they consider to be most likely. One reason for doing so is to assess whether those advisers reached similar conclusions about the likelihood of nuclear use for different reasons. Burying such disagreements can be a natural bureaucratic inclination, but their existence may suggest there is more uncertainty around the consensus probability estimate than is immediately apparent. To be sure, this procedure is not a panacea; groupthink may result in advisers' focusing on the same escalation pathway (and for this reason, decisionmakers could try, to the extent possible, to consult with advisers individually). Even so, canvassing advisers about the most dangerous escalation pathways should help decisionmakers develop more robust policies by not fixating on any one of them.

The decisionmakers responsible for navigating any future deep crisis or conventional conflict between nuclear-armed states will not have the luxury of calling for more research on escalation before making decisions. They will have to act—or not act—based on the understanding of escalation that they and their advisers have, not on the one they would ideally want. Given the extent of the uncertainty about escalation, forecasting will not yield reliable estimates of how the likelihood of escalation will change under different courses of action, let alone tell the decisionmakers how to act. But, if carefully and modestly applied, it might help leaders better understand the range of the possible and provide a useful input to the decisionmaking process.

Appendix 1: Workshop Participants

Three experts included here participated in a practice workshop we hosted to refine the exercise design. Their data is not captured in the analysis presented throughout this paper.

Kil Joo Ban, Korea University

Elaine Bunn, Expert Consultant

Toby Dalton, Carnegie Endowment for International Peace

Thomas Fingar, Stanford University

Markus Garlauskas, Atlantic Council

Matt Gentzel, Longview Philanthropy

Bonnie Glaser, German Marshall Fund

Sam Glover, Forecasting Research Institute

Bethany Goldblum, University of California, Berkeley

Nigel Gould-Davies, International Institute for Strategic Studies

Finn Hambly, Swift Centre

Dominic Johnson, Oxford University

Shashank Joshi, *The Economist*

Isabel Juniewicz, Open Philanthropy

Lami Kim, Daniel K. Inouye Asia-Pacific Center for Security Studies

Arie Kruglanski, University of Maryland

Jeffrey Lewis, James Martin Center for Nonproliferation Studies

Narushige Michishita, National Graduate Institute for Policy Studies (GRIPS)

Rachel Minyoung Lee, Stimson Center

Kjirste Morrell, Good Judgement, Inc.

Anna Nettleship, King's College London

Hanna Notte, James Martin Center for Nonproliferation Studies

Reid Pauly, Brown University

Andreas Persbo, Open Nuclear Network

Andrew Reddie, Berkeley Risk and Security Lab

Brad Roberts, Center for Global Security Research at Lawrence Livermore National Laboratory

Josh Rosenberg, Forecasting Research Institute

Jacquelyn Schneider, Stanford University

Philipp Schoenegger, London School of Economics

Peter Scoblic, New America

Graham Stacey, European Leadership Network

Chris Steinitz, CNA

Lauren Sukin, London School of Economics

Jessica Taylor, Princeton University

Courtney Tee, Global Shield

Bruno Tertrais, Fondation pour la Recherche Stratégique

Jenny Town, Stimson Center

Tong Zhao, Carnegie Endowment for International Peace

Katarzyna Zysk, Norwegian Institute for Defence Studies

Appendix 2: Example Assumptions and Scenario

These example assumptions and scenario are included here exactly as they were provided to participants at the North Korea workshop, which we held in May 2024. No additional edits or updates have been made.

North Korea Scenario Assumptions

It is February 1, 2029. A new American president has just entered the White House, the South Korean president has been in office since 2027, and Kim Jong Un is still the leader of North Korea. The U.S.-South Korean alliance remains in force and has not undergone major changes since 2024.

U.S. Posture and Capabilities

- Political leaders in the United States continue to emphasize the U.S. "ironclad" commitment to South Korea. The USFK Commander retains war-time operational control of Combined Forces.

- In its declaratory policy, the United States continues to state that "any nuclear attack by North Korea against the United States or its Allies and partners is unacceptable and will result in the end of that regime. There is no scenario in which the Kim regime could employ nuclear weapons and survive."

- The United States has delivered on its promise to "enhance the regular visibility" of its strategic assets on and around the peninsula, as laid out in the 2023 Washington Declaration, and has increased the number of these visits in recent months. Annual military exercises have evolved to include nuclear response planning, focused on coordination of South Korean strategic and U.S. nuclear capabilities in an escalating crisis. The United States has not redeployed nuclear weapons to the Korean peninsula.

- POTUS and POTROK have established a secure communication channel for prompt consultations in a crisis that is regularly exercised.

- The United States and South Korea enjoy conventional superiority on the Korean peninsula. These capabilities have continued to evolve since 2024.

- Key U.S. capabilities:
  - A multi-layered missile defense system deployed in South Korea, designed to protect high-value targets against aircraft and short-range missiles (Patriot-2 and Patriot-3), to provide area defense against short- and medium-range missiles (Terminal High Altitude Area Defense (THAAD)), and to provide sea-based defenses against regional ballistic missile threats (Aegis)
  - A real-time mechanism for sharing North Korean missile warning data with South Korea and Japan
  - Hundreds of precision-strike missiles—including submarine-launched cruise missiles, surface-to-surface missiles (ATACMS on HIMARS), multiple rocket launchers (M270), and mid-range ground-launched missile launchers (Typhon)— that allow for deep, short-notice conventional strikes into North Korea

South Korean Posture and Capabilities

- Alongside its alliance with the United States and their combined defense posture, South Korea places its so-called three-axis system at the center of its deterrence and defense strategy. The system has three components: (1) Kill Chain for preempting attacks, (2) Korea Air and Missile Defense (KAMD) for intercepting attacks, and (3) Korea Massive Punishment and Retaliation (KMPR) for retaliating after an attack.
- South Korea's Strategic Command (ROKSTRATCOM), which manages the three-axis system, has been operational since 2024. There has been some coordination between ROKSTRATCOM and U.S.-ROK Combined Forces Command (CFC). However, unlike other parts of the South Korean military, ROKSTRATCOM would not come under the operational command of CFC in a conflict.
- Key South Korean capabilities:
  - Limited independent satellite surveillance capabilities
  - A multi-layered missile defense system designed to protect high-value targets against aircraft and short-range missiles (Patriot-2 and Patriot-3; Cheongung) and to provide area defense against short- and medium-range missiles (L-SAM)

- Thousands of precision-strike missiles—including multiple rocket launchers (K239 Cheonmu), ground-launched ballistic and cruise missiles (ATACMS, and multiple Hyunmoo variants, including with penetrator payloads), air-dropped bunker buster bombs (GBU-28), and air-launched cruise missiles (AGM-84H/K SLAM-ER and KEPD 350 Taurus ALCM)—that can hit targets anywhere in North Korea

- 50 F-35A stealth fighter jets (and various non-stealthy fighters)

North Korean Posture and Capabilities

- North Korea continues to embrace the nuclear strategy outlined in the 2022 Nuclear Forces Policy Law. It has identified two core roles for its nuclear forces: 1) deter attacks "seriously threatening the security of the country and the people," and 2) use nuclear weapons to repel attacks if deterrence fails.

- North Korea has emphasized Pyongyang's right to use nuclear weapons preemptively and has reiterated that "a nuclear strike shall be launched automatically and immediately" according to an "operational plan decided in advance" should Kim's command and control be threatened by an adversary's attack.

- At the 9th Worker's Party Congress in 2026, Kim set out a new five-year military modernization and expansion agenda. Key goals include launching more reconnaissance satellites, improving maneuverable reentry vehicle technology, and ensuring the survivability and effectiveness of the nuclear arsenal. An ongoing scientific exchange between North Korea and Russia, which began in 2023, has helped North Korea to advance these goals, especially by refining space launch capabilities and accessing the materials needed to scale up solid-fuel missile production.

- Key North Korean capabilities:

  - 90-120 nuclear warheads, including high-yield thermonuclear and low-yield tactical warheads

  - Hundreds of ground-launched regional ballistic missiles (solid and liquid-fueled, assumed assigned to both conventional and nuclear missions) with diverse basing modes, including rail-mobile launchers, fixed silos, TELs, and lake-submerged launchers

  - Tens of intercontinental ballistic missiles (solid and liquid-fueled, all assumed assigned exclusively to nuclear missions), including some with multiple independent reentry vehicles

- A small force of regional submarine-launched ballistic and cruise missiles (all assumed assigned exclusively to nuclear missions), deployable on the country's three ballistic missile submarines (SSBs)

- A small force of ground-launched nuclear-capable cruise missiles

North Korea Scenario 1: Nuclear Redeployment

The United States and South Korea release a joint press statement following the thirteenth Nuclear Consultative Group (NCG) principals meeting, announcing that:

> At the direction of the Presidents of the United States and the Republic of Korea, the alliance will begin preparations to deploy U.S. nonstrategic nuclear warheads to the Republic of Korea. This deployment, which is strictly defensive, is intended to enhance deterrence.
>
> The warheads will remain in U.S. custody and control in full compliance with the Treaty on the Non-Proliferation of Nuclear Weapons. Certified South Korean dual-capable aircraft will be made available for nuclear roles and South Korean personnel will be trained accordingly.
>
> As a first step, in the coming days, the alliance will begin constructing facilities at Kunsan Air Base capable of safely storing the warheads.

Following the announcement, a Korean Central News Agency (KCNA) statement warns that North Korea "will not tolerate U.S. plans to arm South Korea with nuclear weapons," and that "if the American imperialists and their illegitimate lackeys try to bring nuclear war to our peninsula, we will have no choice but to strike first in self-defense."

Two months later, the United States and South Korea commence their annual Ulchi Freedom Shield (UFS) exercise. The exercise includes aerial drills over the East Sea, close to North Korean airspace. The aircraft involved include South Korean F-35As stationed at Kunsan Air Base.

Partway through the drills, North Korea fires surface-to-air missiles (SAMs) at allied aircraft participating in the drills. The aircraft evade the missiles. A joint U.S.-South Korean intelligence assessment in the immediate aftermath of the incident concludes with high confidence that the SAMs were launched with the intention to shoot down aircraft and not as a warning shot. The U.S. and South Korean presidents consult and order a retaliatory strike on the SAM battery and radars that carried out the launch. The allied strike, conducted jointly by U.S. and South Korean aircraft, destroys those assets, killing 12 North Korean soldiers in the process.

A Korean Central Television (KCTV) broadcast characterizes the incident as "an unjust response to the brave defense of our sovereign airspace," and accuses the United States of "irresponsibly equipping feeble South Korean pilots with nuclear power." The broadcast goes on to say that "Marshal Kim has vowed to respond in an appropriate manner without delay to defend the honor of our fallen comrades."

Days later, North Korea launches a salvo of ten conventionally armed ballistic and cruise missiles at Kunsan Air Base. U.S. missile defense systems deployed in South Korea intercept six missiles. Four missiles strike the base, temporarily disabling three aircraft hangars and causing significant damage at the construction site of an underground storage vault for nuclear warheads. The strike kills eight South Korean and five U.S. military personnel, and injures an additional 25.

About the Authors

Jamie Kwong is a fellow in the Nuclear Policy Program at the Carnegie Endowment for International Peace. Her research focuses on nonproliferation issues, the Korean Peninsula, and multilateral regimes, including the P5 Process and the Nuclear Non-Proliferation Treaty. She has also conducted novel research on the climate change-nuclear weapons nexus and authored the Carnegie paper, "How Climate Change Challenges the U.S. Nuclear Deterrent."

Anna Bartoux is a research assistant in the Carnegie Nuclear Policy Program. She was previously a James C. Gaither Junior Fellow in the Carnegie Nuclear Policy Program. Before joining Carnegie, Anna studied Political Science at Columbia University.

James Acton holds the Jessica T. Mathews Chair and is co-director of the Nuclear Policy Program at the Carnegie Endowment for International Peace. A physicist by training, Acton is currently writing a book on the nuclear escalation risks of advanced nonnuclear weapons and how to mitigate them. His work on this subject includes the *International Security* article "Escalation through Entanglement" and the Carnegie report, *Is It a Nuke?*

Acknowledgements

This work was made possible by generous support from Effective Ventures Foundation USA and Effective Ventures Foundation (UK). We thank Longview Philanthropy for facilitating funding. We also thank the reviewers for their invaluable comments. Naturally, the authors take full responsibility for this paper's contents.

Notes

1 James Acton, "How I Learned to Stop Worrying and (Sort Of) Love Nuclear Forecasting," Metaculus, April 13, 2023, https://www.metaculus.com/notebooks/15906/how-i-learned-to-stop-worrying-and-sort-of-love-nuclear-forecasting/.

2 For other examples, see Patricia Lewis, Sasan Aghlani, Benoît Pelopidas, and Heather Williams, "12 Times We Came Close to Using Nuclear Weapons," Chatham House, February 1, 2021, https://www.chathamhouse.org/2016/07/12-times-we-came-close-using-nuclear-weapons; and Richard K. Betts, *Nuclear Blackmail and Nuclear Balance* (Washington, DC: The Brookings Institution, 1987), especially 82–131.

3 Bob Woodward, *War* (New York: Simon & Schuster, 2024), 151.

4 David Sanger, "Biden's Armageddon Moment: When Nuclear Detonation Seemed Possible in Ukraine," *The New York Times*, March 9, 2024, https://www.nytimes.com/2024/03/09/us/politics/biden-nuclear-russia-ukraine.html; Lau Stuart, "China's Xi Warns Putin Not to Use Nuclear Arms in Ukraine," Politico, November 4, 2022, https://www.politico.eu/article/china-xi-jinping-warns-vladimir-putin-not-to-use-nuclear-arms-in-ukraine-olaf-scholz-germany-peace-talks/; and Vivian Salama and Michael Gordon, "Senior White House Official Involved in Undisclosed Talks With Top Putin Aides," *Wall Street Journal*, November 7, 2022, https://www.wsj.com/articles/senior-white-house-official-involved-in-undisclosed-talks-with-top-putin-aides-11667768988.

5 Demetri Sevastopulo, "Antony Blinken: 'China Has Been Trying to Have it Both Ways,'" *Financial Times*, January 3, 2025, https://www.ft.com/content/25798b9f-1ad9-4f7f-ab9e-d6f36bbe3edf.

6 James G. Blight and David A Welch, *On the Brink: Americans and Soviets Reexamine the Cuban Missile Crisis* (New York: Hill and Wang, 1989), 214–215.

7 See, for example, Bridget Williams et al., *Can Humanity Achieve a Century of Nuclear Peace? Expert Forecasts of Nuclear Risk*, FRI Working Paper #4 (Forecasting Research Institute and Open Nuclear Network, October 29, 2024), https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/672541e94c430d6f5583a2f9/1730494967571/NuclearRisk.pdf.

8 NunoSempere, Misha_Yagudin, and elifland, "Samotsvety Nuclear Risk Forecasts — March 2022," Effective Altruism Forum, March 10, 2022, https://forum.effectivealtruism.org/posts/KRFXjCqqfGQAYirm5/samotsvety-nuclear-risk-forecasts-march-2022#Methodology. For a thoughtful critique of this assessment, see J. Peter Scoblic, "Nuclear Expert Comment on Samotsvety Nuclear Risk Forecast," Effective Altruism Forum, March 26, 2022, https://forum.effectivealtruism.org/posts/W8dpCJGkwrwn7BfLk/nuclear-expert-comment-on-samotsvety-nuclear-risk-forecast-2.

9 If some event has a probability of p_A if political party A is in power and a probability of p_B otherwise, then the probability of its occurrence is $q_A p_A+(1-q_A)p_B$, where q_A is the probability of party A's holding power.

10 Available at https://carnegie-production-assets.s3.amazonaws.com/static/files/Nuclear_Escalation_Workshop.pdf.

11 For example, the *unconditional* probability of drawing an ace from a normal deck of playing cards is about 8 percent (4/52). The *conditional* probability of drawing an ace from a deck of cards given that all number cards have been removed is 25 percent (4/16).

12 Operationally, participants completed this exercise by writing their probability estimates by hand on a printed copy of figure 1.

13 Forrest E. Morgan, Karl P. Mueller, Evan S. Mederios, Keven L. Pollpeter, and Roger Cliff, *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, CA: RAND Corporation, 2008), xi, https://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG614.pdf. See also Thomas C. Schelling, *Arms and Influence* (New Haven, CT: Yale University Press, 1966), 126 –189; and Richard Smoke, *War: Controlling Escalation* (Cambridge, MA: Harvard University Press, 1977), 13–18.

14 By this definition, a state's use of nuclear weapons on its own territory to forestall advancing forces would not constitute nuclear use—a potential bug, albeit one that did not come up at all in any of the workshops.

15 Jeffrey A. Friedman, Joshua D. Baker, Barbara A. Mellers, Philip E. Tetlock, and Richard Zeckhauser, "The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament," *International Studies Quarterly* 62, no, 2 (June 2018): 419, https://doi.org/10.1093/isq/sqx078.

16 Of course, skipped thresholds might subsequently be crossed. For example, if a crisis escalated straight from an adversary nuclear alert (threshold 2) to U.S. nuclear use (threshold 4), it would be entirely possible for the adversary to then use nuclear weapons in the region (threshold 3). However, for this exercise, we defined the concept of an escalation pathway in terms of the highest level of escalation reached at a given time and instructed participants not to add "backward" arrows to figure 1.

17 One plausible—but incorrect—explanation is that, because of the complexity of the exercise, participants were unwilling to assign probabilities to the branch arrows in figure 1. In fact, among participants who estimated a probability for at least one branch arrow, there was no correlation between the number of branch arrows with assigned probability estimates and the probability of escalation's occurring along the principal pathway. (Not assigning probability estimates to any of the branch arrows leads 0-1-2-3-4-5 to be the only possible escalation pathway.)

18 That said, the median of these estimates was almost identical to the arithmetic mean so this choice made no material difference.

19 See, for example, Williams et al., "Can Humanity Achieve a Century of Nuclear Peace?"

20 James G. March and Zur Shapira, "Managerial Perspectives on Risk and Risk Taking," *Management Science* 33, no. 11 (November 1987), https://www.jstor.org/stable/2631920.

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 170 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Nuclear Policy Program

The Nuclear Policy Program aims to reduce the risk of nuclear war. Our experts diagnose acute risks stemming from technical and geopolitical developments, generate pragmatic solutions, and use our global network to advance risk-reduction policies. Our work covers deterrence, disarmament, arms control, nonproliferation, and nuclear energy.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org