Issue Brief

Beyond P(doom) for AI Risk

Quantifying Uncertainty Without Probability

Author
Andrew J. Lohn

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

May 2026

Executive Summary

Many observers wonder about the potential for artificial intelligence to cause catastrophic risks, but fortunately, there is little empirical evidence on which to base those assessments. Absent such evidence, experts often use their best guesses to estimate the probability of an AI-induced catastrophe or apocalypse (i.e., p-doom). Although subjective expert assessment may be the best evidence available, policymakers and risk analysts are not restricted to asking for probabilities. This brief promotes additional tools for handling uncertainty in AI risk assessments.

Imagine you are asked to roll a six-sided die but you have only seen three sides; one side has a star etched on it, two sides are blank, and the other three are unknown. Predicting the outcome involves part randomness and part ignorance. If you are asked to give the probability of a star, you might note that one of the three sides you've seen has a star and answer $1/3$. Asked how confident to be that a star will come up, there is only one side that you know has a star, so $1/6$ would be a reasonable answer. Asked whether a star could come up, you might note that four sides could have a star, so $4/6$ would be a reasonable answer. These questions appear similar, but their differences are important if you are a decision-maker who cares about stars.

In AI risk, rather than in dice rolls, ignorance is the dominant form of uncertainty, not randomness, so the best techniques are not always probabilistic. There are alternative mathematical techniques that are just as rigorous as probability. They also use familiar terms from common discourse, such as Belief and Plausibility, allowing them to easily become part of popular AI risk vernacular and to be communicated to decision makers.

The way to think of the mathematical term Belief is that it expresses how confident one can be based on the evidence. For instance, the evidence allows us to have a 1/6 degree of belief that the die will come up stars. Plausibility expresses what is left after removing the counter-evidence. Two of the six sides cannot be stars, so the Plausibility of stars is 4/6. The gap between Belief and Plausibility is due to ignorance. Without ignorance, Belief and Plausibility become the same number, equal to probability.

This issue brief explains why analysts and decision-makers need alternatives to probability for handling the uncertainty in AI risk. It explains Belief, Plausibility, and how they relate to probability in an intuitively accessible way. And it demonstrates how to calculate Belief and Plausibility in the context of expert assessments of AI risk.

At a high level, enacting the change sought by this brief is easy. Policymakers only need to add two additional questions when discussing AI risks. The first is either, how certain are you that this risk will occur, or even better, how strong is the evidence supporting this hypothetical outcome? The second is, how certain are you that this risk will *not* occur, or how strong is the evidence *against* this hypothetical outcome?

Asking those two additional questions will force analysts to confront their sources of uncertainty more directly and drive analysts to expand their risk analysis toolbox. Answering those questions, and communicating those answers, is also a low lift because the analytical techniques already exist and because the vocabulary is already familiar. This brief provides an introduction to those techniques and vocabulary.

Table of Contents

Introduction

Experts, policymakers, and citizens around the world debate the existential risks of AI with astonishingly little agreement.[1] One study found that estimates for the likelihood of AI-induced existential catastrophes differed by 250 times between a group of "skeptics" and a "concerned" group (0.001 and 0.25, respectively).[2] Even within groups, such as among AI experts, the variation can be jarring.[3]

This report aims to familiarize analysts and decision-makers with the tools and vocabulary to handle this uncertainty more explicitly. It also suggests that some of this disagreement may be due to imprecise terminology rather than wholly from substantive differences in views. This brief aims to provide conceptual clarity to resolve disagreements where discussants talk past each other in describing their expectations about the future of AI risk. That clarity may help find common ground and perhaps help to pinpoint the sources of the disagreements or identify the data that would be needed to resolve them.

Probability Is Not the Only Option

Existential AI risk questions are usually phrased along the lines of, what is your estimate for p(doom) (i.e., probability of AI-induced apocalypse), or what do you think are the odds of a major AI catastrophe in the next five years? Setting aside the ambiguity about how much destruction is required to be existential or catastrophic, the deeper problem is that probability is poorly suited to answering these types of questions. The upside is that there are additional options, some of which are based on terms that are already familiar to both experts and laypeople even if the precise quantitative definitions are not.

This section discusses why probability is ill-suited to the topic of existential risks from AI. The next sections introduce alternatives to probability—namely, Belief and Plausibility—and explain their advantages for handling various types of uncertainty. The final section illustrates how to combine expert assessments of risk using these concepts to better understand where disagreements in risk assessments truly lie.

Aleatoric and Epistemic Uncertainty

Famously, there are known unknowns and unknown unknowns.[4] More technically, risk analysts separate uncertainty into *aleatoric* and *epistemic* categories. Aleatoric uncertainty deals with variability or indeterminacy that naturally occurs in a system. Epistemic uncertainty deals with a lack of knowledge.[5] Put another way, there is randomness and there is ignorance.[6] Aleatoric uncertainty is not knowing which face of the die will come up. Epistemic uncertainty is not knowing what is written on those faces. Probability is excellent for calculating which face will come up, but it is not ideal for estimating what will be written on it. With epistemic uncertainty, it might not even be clear what a stated probability means or how to calculate it.

As a result, probability is often an inappropriate framing of AI risk that can skew perspectives in ways that harmfully detract from effective discourse and decision-making. In writing about risk in general (not specifically about AI), Terje Aven, the former President of the Society for Risk Analysis, said that "some of the current perspectives . . . are simply misguiding the decision-maker in many cases." He has also said that "probability has to be removed from the definition of risk and the natural replacement is uncertainty."[7]

Imagine a six-sided die where you have only seen three sides: one side has a star etched on it, two sides are blank, and the other three are unknown. You may want to

know how likely it is for a star to come up. Or instead, you may want to know how confidently to believe that a star will come up. Or you may want to know how plausible it is that a star could come up. Those appear to be similar questions but they result in different answers and, in a policy context, have different implications.

You can have $1/6$ (~17%) confidence that a star will come up on the next roll because you know that there is at least one star. That does not make you 83% confident that a star will not come up, because several more sides could also have stars. But knowing that two sides do not have a star is useful information. It makes you $4/6$ (~67%) confident that a star could plausibly come up. That would require all three of the unknown sides to have stars, but that is plausible. To estimate likelihood, you could use the information that you do have. You could say that, because there are stars on one-third of the sides you have seen, there is a $1/3$ (~33%) chance of rolling a star.

These answers are different because there is an underlying mix of aleatoric and epistemic uncertainty. The roll of the die is random, but you are also ignorant about what is written on three of the faces that could come up. Most real-world uncertainties are a mix of randomness and ignorance. In cases that are well-understood or that have plenty of evidence to draw from, randomness can dominate. As an example, cyber intrusions are common enough to collect statistics and have causal chains of events that can be delineated and reasoned about. As a result, the techniques for handling aleatoric uncertainty are often appropriate even though anticipating cyberattacks still involves plenty of ignorance from many sources, including attacker motivations, novel attack methods, or the complexity of network effects.

In AI, where future technology interacts with complex social, economic, or geopolitical systems, there is little data and much is poorly understood. That is especially true of new existential risks, such as from AI, that do not have a history of repeated events to draw from. Ignorance is the dominant form of uncertainty, not randomness. For example, experts may be ignorant about when an AI model would engage in deceptive behavior and how society would respond, despite having evidence suggesting that the models randomly maintain their deceptive behavior 85% of the time.[8] As another example, an analysis may show that AI models advocate for tactical nuclear use in 95% of simulations.[9] But experts remain ignorant about how conflicts and crises will develop, how different today's complex multipolar world is from that study's bipolar dynamics circa 1958–1962, and how much humans will defer to AI suggestions for existential decision-making. Studies such as these two are helping to shift uncertainty from epistemic to aleatoric but the lion's share of the work remains to be done.

Uncertainty about existential risks, especially from AI, is going to remain primarily epistemic, not aleatoric, for some time.

Alternatives to Probability

Fortunately, risk analysis and its toolbox has advanced significantly since de Moivre introduced risk as probability and consequence in 1711.[10] A recent review of relevant risk literature noted that "it is difficult to deal with epistemic uncertainty effectively only through probability theory. Therefore, a series of uncertainty theories have been developed to complement probability theory, which include evidence theory, fuzzy sets, possibility theory, convex models, probability box, etc."[11]

These concepts are not just obscure math. Several have the important benefit of already being part of the common vernacular. That familiarity means that these concepts can be immediately and seamlessly incorporated in AI risk discussions. The math then allows specialists and practitioners to refine those concepts, making the discussions progressively more precise and nuanced while shrinking the bounds of uncertainty.

Belief, Plausibility, and Probability

This section focuses on the familiar terms Belief and Plausibility, and their relation to probability.[12] It provides informal definitions before describing how analysts can use evidence to refine them.

Belief

Belief is the strength of the evidence in favor of a proposition. That is different from the use of "belief" in a religious or faith-based sense. In this technical context, Belief is the degree of certainty in a statement, based on evidence or reason. Notably, that evidence may be subjective, as in the case of expert opinions.

Plausibility

Plausibility is the absence of evidence refuting a proposition. After you remove the degree of Belief that a proposition is not true, its Plausibility remains.

Probability

It is tempting to think of Belief and Plausibility as lower and upper bounds on probability, but that is not always appropriate.[13] Combining conflicting evidence from multiple sources, such as from dissenting expert judgments, can break the simple interpretation of Belief and Plausibility as probability bounds. Belief and Plausibility are adept at handling distinctions between conflicting assessments but in ways that do not always adhere to probability theory. Depending on the amount of conflict among experts and the amount of ignorance, Belief and Plausibility can be bounds for probability. Belief, Plausibility, and probability can even all reduce to the same number. The next section describes how Belief and Plausibility handle conflicting evidence and how they relate to probability.

Structure of Evidence

This section outlines several different ways to classify uncertainty and the several different techniques for handling different types of uncertainty. Understanding the uncertainty is important for making reasoned decisions about it.

Indirect, Subjective, and Conflicting

Belief and Plausibility are most useful when the evidence is not related directly to the question of interest. The most famous example has two witnesses in a courtroom.[14] Each makes conflicting statements and you need to determine the truth. You have no evidence about the crime itself. Instead, you have evidence (subjective assessments) about the reliability of the witnesses.

Imagine that Mrs. Peacock is a witness who is 80% reliable and that she confidently accuses Colonel Mustard and his candlestick. That gives you 80% Belief that Colonel Mustard is the killer, but it does not make you 20% certain of his innocence as it would in basic probability theory where probabilities must sum to one. The unreliable 20% provides no additional information. Although the probability of Colonel Mustard's guilt is actually higher than 80%, there is no basis for determining how much higher. Belief and Plausibility do not require any additional assertions.

Suppose now that Miss Scarlett is 75% reliable and confidently accuses Professor Plum and his wrench. When combining their conflicting testimonies, the jury must recognize that it is not possible for both witnesses to be reliable. The total evidence against Colonel Mustard is the 80% that Mrs. Peacock is reliable times the 25% that Miss Scarlett is not. That is the same answer that probability theory would give if we wrongly ignored that the probability for each defendant should be higher than their accuser's reliability. Belief and Plausibility are built to reason through the remaining cases and to directly consider the conflict among accusations.

In the AI risk context, there are many prominent (i.e., reliable) experts who assert that various AI catastrophes are imminent. There are also many prominent experts who are skeptical of these risks. Both groups might claim to be trustworthy, but they cannot both be correct. The following sections will illustrate how Belief and Plausibility can be used to understand, assess, and calculate uncertainties in the context of AI risk.

Asserting Probabilities

When experts provide subjective probabilities, they are internally weighing the evidence in favor of a proposition and the evidence against it. Their probability reflects a balance between the two. So their answer to the question "What is the probability of an AI catastrophe?" is likely to be higher than their answer to "How certain are you that an AI catastrophe will occur?"

For the die of stars described earlier, the evidence was 17% in favor of stars, but the probability was 33%. Although there was only evidence of one star among the six sides, a reasonable probability estimate would also include the chance that one or more of the unknown sides could have a star. Using a single probability conflates the questions of likelihood, certainty, and plausibility. It is useful to separate those ideas and to be more explicit about the uncertainties.

In one study of AI-induced extinction by the year 2100, the median superforecaster's likelihood was 0.0038.[15] Given the large epistemic uncertainty about existential or catastrophic AI risks, their confidence in catastrophe was presumably much lower. It would be useful to know how much lower.

Confidence Triplets

Rather than using a single probability to represent uncertainty, AI risk analysts can create a triplet of confidence that includes: certainty in catastrophe (C), certainty in no catastrophe (N), and ignorance (I) that all add up to one. That assessment can be written as: [C, N, I]. For the die of stars described earlier, the triplet would be [1/6, 2/6, 3/6].

Experts could be asked to provide that triplet directly or analysts can try to create it from an expert's statement of probability, but creating it from a single probability introduces challenges. There are two bookends on how to do that conversion. Let's say the expert is 20% certain of catastrophe. At one end, the analyst can presume that the expert is partly certain about catastrophe and is ignorant about the rest; their triplet would be [0.2, 0, 0.8]. Alternatively, the analyst can presume that the expert is implying certainty that no catastrophe will occur, for a triplet of [0.2, 0.8, 0].

Even if the expert presumes to have no ignorance, an analyst could choose to adjust the expert assessment to reflect that expert's partial knowledge of the issue. The analyst could attribute a 30% reliability to the expert and adjust the triple of [0.2, 0.8,

0] to [0.2x0.3, 0.8x0.3, 1-0.3] = [0.06, 0.24, 0.7]. The analyst may adjust downward still further if they suspect that the expert is not asserting certainty about catastrophe at all, but is rather asserting a likelihood. In the same way that the confidence in rolling stars was only 17% when the probability was 33%, converting probability estimates for AI risk to certainty should lead to lower numbers.

Confidence Triplets from AI Risk Probabilities

Revisiting the study from the introduction, the "concerned" group's probability of AI-induced extinction by 2100 of 0.25 could correspond to many possible triplets.* For example, both of the following triplets give a probability of 0.25: [0.25, 0.75, 0] or [0.001, 0.003, 0.996]. The former would imply that the expert believes the catastrophe is mostly implausible (Plausibility = 0.25). In that case, with no ignorance, all three of Belief, Plausibility, and probability are 0.25. In the latter triplet, with high ignorance, the expert is claiming that catastrophe is almost certainly plausible (0.997), but that they do not have much evidence to suggest that it will happen (0.001). These are different assertions to a policymaker.

The 0.25 estimate is particularly interesting because it is a very high number for an existential risk, but it is not especially high for Plausibility of an assertion with so much uncertainty. Getting a higher Plausibility while keeping probability at 0.25 requires Belief in catastrophe to be low. Perhaps the experts do mean that there is little evidence to support claims of catastrophe but that their ignorance is high, as in [0.001, 0.003, 0.996]. Or perhaps they mean that catastrophe is both probable and implausible as in [0.25, 0.75, 0]. Or perhaps they are not really expressing probability at all. Perhaps they mean for both Belief and Plausibility to be high, as in [0.25, 0, 0.75], but they use the vocabulary of probability when they mean to refer to Belief.†

Determining an expert's confidence triplet can be done with two questions, and to convert from probability requires one additional question and an assumption. There are three variables to determine: C, N, and I. They have to sum to one, so knowing two is sufficient to calculate the third. For example, given a probability estimate that is an

---

* Probability is calculated from the values C and N alone and does not depend on ignorance (i.e., p=C/(C+N)). From a Bayes' perspective, ignorance is effectively a uniform prior (i.e., 0.5).

† This is common in discourse and is a benefit of Belief and Plausibility that even its critics appreciate. See page 383 of Judea Pearl, "Reasoning With Belief Functions: An Analysis of Compatibility," *International Journal of Approximate Reasoning*, Volume 4, 1990, https://doi.org/10.1016/0888-613X(90)90013-R.

honest probability, rather than being a Belief estimate referred to as probability, then just one more question is needed to calculate the full triplet. That question could ask for Plausibility: "How plausible is AI extinction by 2100?" Or it could ask for Ignorance: "What fraction of the information needed to make this assessment does the expert know?"

Sources of Disagreements

Focusing on sources of disagreement, it is also at least mathematically possible for the skeptics and concerned groups from that study to have the exact same Belief about catastrophe despite probabilities that are 250 times different. The skeptics could have a triplet of [0.0005, 0.4995, 0.5] to get their 0.001 likelihood, and the concerned could have a triplet of [0.0005, 0.0015, 0.998] for their 0.25 likelihood. With those triplets, the concerned group would view catastrophe as both more likely and more plausible despite equivalent levels of Belief. In that case, the disagreement would not be about the case for catastrophe. The disagreement would focus on the case against catastrophe and how much the experts do not know (i.e., ignorance).

Without having asked the experts more directly about their uncertainties, an analyst cannot know which they meant to express, but there is some qualitative information about these two particular groups from a corresponding study.[16] Members of the concerned group were swayed by the precedent that higher intelligences had previously caused extinction of lesser intelligences, whereas the skeptics felt that sweeping changes tend to be slow. Based on those arguments, the two groups would have different levels of Belief. The concerned group would, unsurprisingly, have high values for C, while the skeptics would have high values for N.

More interestingly, the concerned group also was "more willing to place weight on theoretical arguments with multiple steps of logic, while the skeptics tended to doubt the usefulness of such arguments." That implies that the two groups handled ignorance differently, but there are too many ways to interpret that qualitative statement without having asked an additional question that could be used to quantify their ignorance, Belief, or Plausibility. While we have yet to see Belief and Plausibility explored for AI risk, the authors of the AI risk studies described above have more recently stated that, going forward, more weight should be given to alternative frameworks, including the techniques described in this brief.[17]

Combining Evidence

There are two different objectives for combining expert assessments. One is to add new information when a new independent expert is added. The other is to find consensus among experts who have differing assessments of the same evidence. This section will start by describing techniques for independent experts, then finish with techniques for aggregating assessments of the same evidence.

Let's consider two hypothetical experts, A and B, who are 15% and 20% confident in catastrophe, respectively. Those numbers are very high but make for simple illustration. The first question is what to do with the remaining 85% and 80%. Should it be assigned to ignorance or does it imply that the experts are certain that no catastrophe will occur?

This section will consider three different ways to allocate the remainder: 1) Allocate the remaining 85% and 80% fully to ignorance, 2) allocate those remaining portions fully to certainty that no catastrophe will occur, and 3) have a partial allocation to certainty and ignorance. The subsections below discuss the intuition behind those assignments and the resulting Beliefs and Plausibilities. The calculations themselves are in Appendix A.

Full Ignorance

Assigning the remaining portions all to ignorance gives the triplets [0.15, 0, 0.85] and [0.2, 0, 0.8]. A way to read these triplets is that both experts have nothing to convince them against catastrophe, but they admit that there is much they do not know. As calculated in Appendix A, their combined Belief and Plausibility are both quite high: 0.32 and 1, respectively. Those numbers are high because they each have some reason to suspect catastrophe and no reasoning against it.

Full Certainty

At the other extreme, the experts know everything they need to make the assessments, but the world is still random and could lead to either catastrophe or no catastrophe. That gives the triplets [0.15, 0.85, 0] and [0.2, 0.8, 0]. In this case, there is conflict among the assessments. Handling this conflict is the strength of calculating Belief and Plausibility, but there are different techniques for it.[18] Appendix A shows the calculations for two popular techniques: Dempster's rule and Yager's rule.

Both rules first need to calculate the conflict, which is the total where one expert has some certainty in an outcome that the other has some certainty cannot occur. The first expert assesses 0.15 for catastrophe when the second assesses 0.8 against, which are multiplied to get 0.12. And the first expert assesses 0.85 against catastrophe when the second assesses 0.2 for, which multiplies to 0.17. Adding those two conflicts gives the total conflict: 0.12 + 0.17 = 0.29.

The combination rules also consider the cases where the experts agree. They both agree about catastrophe with assessments of 0.15 and 0.2, which multiply to give 0.03. And they agree about no catastrophe with 0.85 and 0.8, which gives 0.68. Yager's rule simply takes these numbers and then assigns the conflict to ignorance. So, the combined confidence triplet is [0.03, 0.68, 0.29].

Dempster's rule, on the other hand, redistributes the conflict proportionally across the triplet rather than adding the conflict to ignorance. It does that by dividing each of the elements in the confidence triplet by one minus the conflict.* In this case, that divisor would be (1-0.29), so the triplet becomes [0.042, 0.958, 0].

In this case, using Dempster's rule results in no ignorance because neither of the two original assessments had any ignorance. It is like a die where you know all the sides. Without ignorance, Belief and Plausibility reduce to the same number. It is also the same number that results from the probabilistic approach that combines the experts' assessments using Bayes' theorem.[19] In this case, that number is 0.042. That is a small number compared to the experts' individual confidence because, if the experts are both independently providing evidence that no catastrophe will occur, then the odds of catastrophe in their combined assessment should be low. But Yager's rule may be more appropriate if the conflict among experts comes from their partial ignorance.

The Yager's rule combination results in a combined triplet of [0.03, 0.68, 0.29]. So, while Belief stays low at 0.03 (lower than using Dempster's rule), their combined ignorance pushes the Plausibility higher than either of the individual assessments to 0.32.

Mixed Ignorance and Certainty

For less extreme cases, the differences between Dempster's and Yager's rules can be less pronounced. Consider the analyst who assigns only 30% reliability to their two

* The normalization factor for Dempster's rule is 1 – K, where K is the conflict.

experts, resulting in triplets of [0.045, 0.255, 0.7] and [0.06, 0.24, 0.7]. As shown in Appendix A, combining by Dempster's rule (yielding [0.078, 0.419, 0.503]) and Yager's rule ([0.076, 0.408, 0.516]) yield similar results. Belief is 0.078 and 0.076, and Plausibility is 0.58 and 0.59, respectively. The calculated Belief in catastrophe is higher than the certainty of either individual expert because their ignorance is relatively high and their certainty in no catastrophe is relatively low.

Calculating Agreement Rather Than Adding New Evidence

Both Dempster's rule and Yager's rule are meant for combining independent expert assessments, but sometimes the goal of combination is to find the degree of agreement among experts rather than to add new evidence. For example, if two experts review the same articles about AI risk and come to different conclusions, then their assessments are not independent because both are based on the same articles. In that case, the goal is to calculate a single assessment that combines their differing assessments.

In probability theory, that can be done using a technique called logarithmic pooling rather than the standard Bayes' rule.[20] Logarithmic pooling can also be extended for calculating combined Belief and Plausibility, and it can be done in ways that give the same answers as probability theory if there is no ignorance.[21] Although it is not an average, it is conceptually similar to averaging the experts' assessments, but with some properties that make it better suited for combining expert assessments, such as external Bayesianity.*

For the experts above who had triplets of [0.045, 0.255, 0.7] and [0.06, 0.24, 0.7], the combined triplet using logarithmic pooling would be [0.052, 0.248, 0.7]. This calculation is shown in Appendix A.

* Logarithmic pooling retains external Bayesianity, meaning that the order of combining experts does not matter. Averaging would give one answer for combining experts A and B first before combining with C, then a different answer for combining A with C and B with C before combining them together.

Future Work and Limitations

The previous section showed that there are several techniques for combining evidence that can lead to differing results. The appropriate technique for a given problem depends on the independence among the sources of evidence and on the type of conflict among them. It is rare for any technique to be perfectly appropriate in any circumstance, and no technique is always preferred. Further, some of these cases can lead to paradoxical outcomes.[22] We acknowledge that multiple combination techniques that can lead to strange and varied outcomes is an important shortcoming. Nonetheless, we propose that the AI risk community is currently underexploring the implications of how it handles conflicting evidence and that Belief and Plausibility provide mechanics for that exploration. Even the frequent critic of Belief and Plausibility, Judea Pearl, agrees that Belief and Plausibility "offer a rich language for describing the evidence gathered, highly compatible with the way people summarize observations."[23]

Apart from the mechanics of calculations, the primary challenge in eliciting accurate probabilities still remains when eliciting Belief and Plausibility—they are both attempts to quantify the unknowable. Just as how policymakers should be skeptical of expert opinions of the likelihood of AI catastrophes, policymakers should be skeptical of expert opinions on the strength of evidence for and against those same outcomes. There is some reason to expect that Belief and Plausibility may be less accurate than probability because analysts have spent more time practicing and calibrating for probability estimation. There is also some reason to expect that Belief and Plausibility may be more accurate than probability if, as it can be for witness testimony, it is easier to assess expert reliability than it is to assess the truth of their claims directly. We hope that there continue to be too few AI catastrophes or existential events to empirically determine which approach is more accurate. Ultimately though, accuracy is not the primary reason to move away from probability. The objective is to better conceptualize and communicate risk, not just to measure it.

Point estimates of probability do a minimal job of conceptualizing and communicating risk. One step forward from there uses probability distributions rather than point estimates. That has the advantage of being simple and familiar but has the disadvantage of enforcing structure that is rarely justified, such as by assuming a uniform or Gaussian distribution or by requiring that the weight of all theorized outcomes sum to one. More broadly though, the familiarity that is its main advantage is also a shortcoming. Contemporary risk analysis has expanded beyond probability and

AI risk deserves a large body of risk analysts who are familiar with more of those tools. Belief and Plausibility offer a simple departure point for policymakers and AI risk analysts to explore the wider set of tools. Ultimately, another tool, or probability, may be the best option. Our hope with this paper is that the tools will be chosen based on their merits for the case at hand rather than always defaulting to probability merely because of lack of exposure to non-probabilistic techniques.

Conclusion

AI risk analysts, policymakers, and the general public all want to know "How worried should we be about an AI catastrophe?" The question that they often ask instead is "How likely is an AI catastrophe?" But it may be more appropriate and informative to ask "How certain are you that catastrophe will occur?" or "How plausible is an AI catastrophe?" Those are different questions that result in different answers that require different vocabularies and different mathematical techniques.

We recommend that policymakers continue to ask about the probability of worrying events but that they pose two additional questions to their experts and analysts. The first is "How certain are you that this risk will occur?" or "How strong is the evidence supporting this hypothetical outcome?" The second is "How certain are you that this risk will not occur?" or "How strong is the evidence against this hypothetical outcome?"

Those additional questions will force policymakers, experts, and analysts to grapple with the epistemic uncertainty that dominates AI risk more directly, and to avoid conflating epistemic and aleatoric uncertainties the way point estimates and even probability distributions do. We anticipate that those additional questions will help resolve, or at least clarify, the large discrepancies that exist today between expert probability judgments. We also anticipate that making epistemic uncertainty more explicit will lead to more accurate risk assessment by driving efforts to progressively convert epistemic uncertainties into aleatoric uncertainties, then shrinking those aleatoric variances.

The additional questions will also force analysts to expand their toolkit and provide more informative, well-reasoned, and intellectually humble answers without downplaying the risks. In fact, the high degrees of ignorance around AI risks may lead to alarmingly large plausibility numbers that motivate more aggressive policy action.

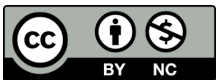
The burden of this recommendation is low. It is simply to ask two questions. The vocabulary of Belief and Plausibility is already familiar and the mathematical techniques already exist. They can be added seamlessly to risk analysis and discourse. Adding them will not magically resolve all of the uncertainty, but it will help to capture the uncertainty in ways that are easier to understand, communicate, and assess. It is well worth it if it can help bring AI risk analysis out of the 1700s and incorporate present-day tools, techniques, and best practices.

Author

Andrew J. Lohn is a senior fellow working on the CyberAI Project at the CSET.

Acknowledgments

The author would like to thank Igor Mikolic-Torreira, John Bansemer, Katherine Quinn, Emmy Probasco, Josh Goldstein, and Owen Daniels for useful feedback on earlier drafts.



Document Identifier: doi: 10.51593/20260001

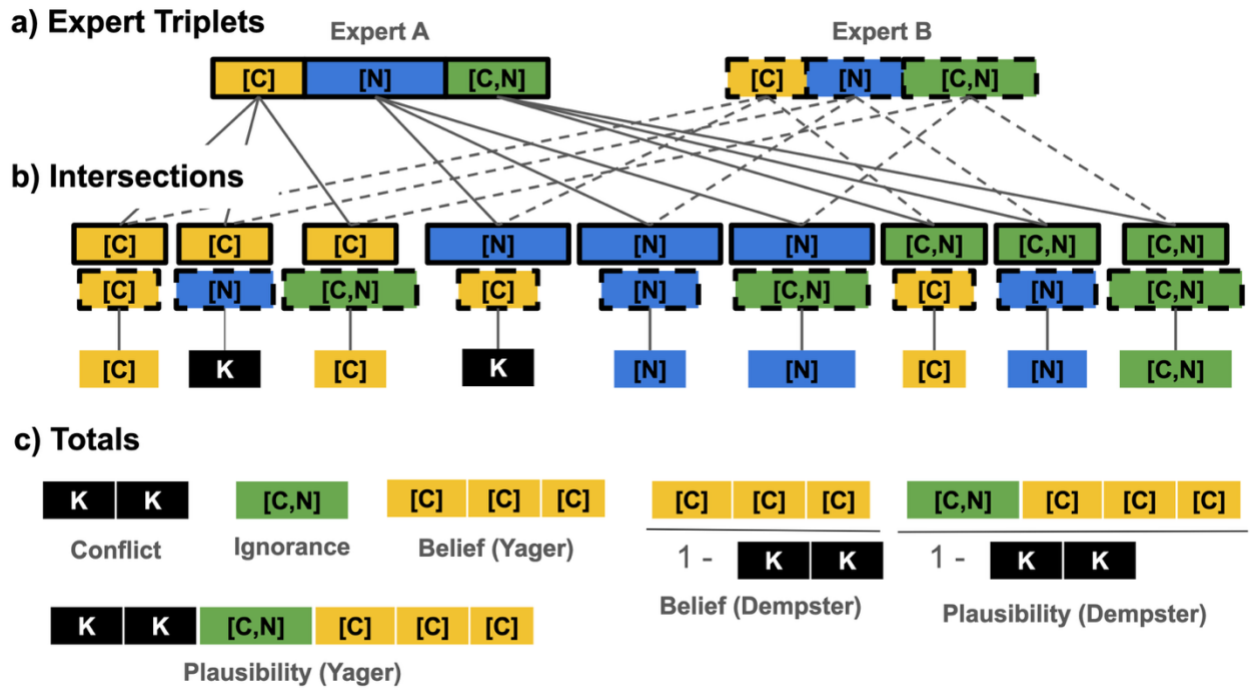
Appendix A: Combining Experts to Calculate Belief and Plausibility

This is not meant to be a tutorial on Dempster-Shafer theory. This appendix is merely meant to be an illustrative walk-through that is as easy to follow as possible. To meet that need, it forgoes definitional rigor and uses minimal mathematical notation. We also use numbers that are easy to work with. The numbers are not meant to reflect any actual expert assessment.

Each expert can have some certainty of catastrophe C, some certainty of no catastrophe (N), and some degree of ignorance. Ignorance is more precisely represented as both C and N together: {C,N}. The goal of combination is to tally how much their assessments overlap with one another. The various intersections can be mapped out in a table, but the first step is to revisit the expert assessments.

Figure A.1 illustrates how the confidence triplets combine for two experts in the figure below and walk through the calculations for the three scenarios described in the main text.

Figure A.1: Combining the confidence triplets of two experts



Full Ignorance

In the example, expert A is 15% certain of catastrophe. Assigning the remaining 85% to ignorance results in a triplet [0.15, 0, 0.85]. Doing the same for expert B gives [0.2, 0, 0.8].

The table calculating the degree of overlap (intersection) between the two assessments is shown below:

Table A.1: The terms used to combine expert assessments by Yager's and Dempster's methods with full ignorance

Expert A	Expert B	Intersection	Total
{C}: 0.15	{C}: 0.2	{C}	0.15 x 0.2 = 0.03
{C}: 0.15	{N}: 0	-	0.15 x 0 = 0
{C}: 0.15	{C,N}: 0.8	{C}	0.15 x 0.8 = 0.12
{N}: 0	{C}: 0.2	-	0 x 0.2 = 0
{N}: 0	{N}: 0	{N}	0 x 0 = 0
{N}: 0	{C,N}: 0.8	{N}	0 x 0.8 = 0
{C,N}: 0.85	{C}: 0.2	{C}	0.85 x 0.2 = 0.17
{C,N}: 0.85	{N}: 0	{N}	0.85 x 0 = 0
{C,N}: 0.85	{C,N}: 0.8	{C,N}	0.85 x 0.8 = 0.68

There are two rows with no intersection between the set {C} and the set {N}. In this case, both of those rows where there is no intersection have 0 mass, so there is no conflict (K).

$$K = 0 + 0 = 0$$

There is also a row with multiple intersections. The total from that one row is called the ignorance:

$$Ignorance = 0.68$$

To calculate Belief, calculate the sum of masses where the intersection is exactly {C}: 0.03 + 0.12 + 0.17 = 0.32

That is the degree of Belief according to Yager's rule.

$$Belief_{Yager} = 0.32$$

Using Dempster's rule, the degree of Belief is that sum above divided by (1 – K). Since K = 0, the Belief is the same.

$$Belief_{Dempster} = \frac{0.32}{1-0} = 0.32$$

To calculate Plausibility, calculate the sum of masses where the intersection includes {C}. That calculation is different between Dempster's rule and Yager's rule. That is because Yager's rule converts the conflict into ignorance. Yager's rule includes both the ignorance and the conflict in calculating Plausibility, whereas Dempster's rule only includes the ignorance.

$$Plausibility_{Yager} = 0.003 + 0.12 + 0.17 + 0.68 + 0 + 0 = 1$$

$$Plausibilty_{Dempster} = \frac{0.03 + 0.12 + 0.17 + 0.68}{1-0} = 1$$

Full Certainty

In the example, expert A is 15% certain of catastrophe. Assigning the remaining 85% to certainty of no catastrophe results in a triplet [0.15, 0.85, 0]. Doing the same for expert B gives [0.2, 0.85, 0].

The table calculating the degree of overlap (intersection) between the two assessments is shown below:

Table A.2: The terms used to combine expert assessments by Yager's and Dempster's methods with full certainty.

Expert A	Expert B	Intersection	Total
{C}: 0.15	{C}: 0.2	{C}	0.15 x 0.2 = 0.03
{C}: 0.15	{N}: 0.8	-	0.15 x 0.8 = 0.12
{C}: 0.15	{C,N}: 0	{C}	0.15 x 0 = 0
{N}: 0.85	{C}: 0.2	-	0.85 x 0.2 = 0.17
{N}: 0.85	{N}: 0.8	{N}	0.85 x 0.8 = 0.68
{N}: 0.85	{C,N}: 0	{N}	0.85 x 0 = 0
{C,N}: 0	{C}: 0.2	{C}	0 x 0.2 = 0
{C,N}: 0	{N}: 0.8	{N}	0 x 0.8 = 0
{C,N}: 0	{C,N}: 0	{C,N}	0 x 0 = 0

There are two rows with no intersection but they now have masses of 0.12 and 0.17, so there is conflict (K).

$$K = 0.12 + 0.17 = 0.29$$

There is also a row with multiple intersections but that ignorance is empty.

$$Ignorance = 0$$

To calculate Belief, calculate the sum of masses where the intersection is exactly {C}: 0.03 + 0 + 0 = 0.03

That is the degree of Belief according to Yager's rule.

$$Belief_{Yager} = 0.03$$

Using Dempster's rule, the degree of Belief is that sum above divided by (1 – K).

$$Belief_{Dempster} = \frac{0.03}{1 - 0.29} = 0.042$$

To calculate Plausibility, calculate the sum of masses where the intersection includes {C}. With Dempster's rule there is no ignorance so it is the same calculation as Belief.

$$Plausibilty_{Dempster} = \frac{0.03 + 0 + 0 + 0}{1 - 0.29} = 0.042$$

Yager's rule converts the conflict into ignorance.

$$Plausibility_{Yager} = 0.003 + 0 + 0 + 0.29 = 0.32$$

Partial Certainty and Partial Ignorance

In the example, expert A is 15% certain of catastrophe. Assigning the remaining 85% to certainty of no catastrophe but then only being 30% confident in their assessment results in a triplet [0.045, 0.255, 0.7]. Doing the same for expert B gives [0.06, 0.24, 0.7].

The table calculating the degree of overlap (intersection) between the two assessments is shown below:

Table A.3: The terms used to combine expert assessments by Yager's and Dempster's methods with partial certainty and partial ignorance

Expert A	Expert B	Intersection	Total
{C}: 0.045	{C}: 0.06	{C}	0.045 x 0.06 = 0.0027
{C}: 0.045	{N}: 0.24	-	0.045 x 0.24 = 0.0108
{C}: 0.045	{C,N}: 0.7	{C}	0.045 x 0.7 = 0.0315
{N}: 0.255	{C}: 0.06	-	0.255 x 0.06 = 0.0153
{N}: 0.255	{N}: 0.24	{N}	0.255 x 0.24 = 0.0612
{N}: 0.255	{C,N}: 0.7	{N}	0.255 x 0.7 = 0.1785
{C,N}: 0.7	{C}: 0.06	{C}	0.7 x 0.06 = 0.042
{C,N}: 0.7	{N}: 0.24	{N}	0.7 x 0.24 = 0.168
{C,N}: 0.7	{C,N}: 0.7	{C,N}	0.7 x 0.7 = 0.49

There are two rows with no intersection but now have masses of 0.12 and 0.17, so there is conflict (K).

$$K = 0.0108 + 0.0153 = 0.0261$$

There is also a row with multiple intersections but that ignorance is empty.

$Ignorance = 0.49$

To calculate Belief, calculate the sum of masses where the intersection is exactly {C}: 0.0027 + 0.0315 + 0.042 = 0.0762

That is the degree of Belief according to Yager's rule.

$$Belief_{Yager} = 0.0762$$

Using Dempster's rule, the degree of Belief is that sum above divided by (1 – K).

$$Belief_{Dempster} = \frac{0.0762}{1 - 0.0261} = 0.0782$$

To calculate Plausibility, calculate the sum of masses where the intersection includes {C}. With Dempster's rule there is no ignorance so it is the same calculation as Belief.

$$Plausibilty_{Dempster} = \frac{0.0762 + 0.49}{1 - 0.0261} = 0.58$$

Yager's rule converts the conflict into ignorance.

$$Plausibility_{Yager} = 0.0762 + 0.49 + 0.0261 = 0.59$$

Logarithmic Pooling

The combination methods above multiply the masses from each expert. Those are the appropriate approaches if each expert is providing independent assessments and you want to combine their independent contributions. If you want to calculate a combined assessment for their differing interpretations of the evidence, then logarithmic pooling may be more appropriate than the multiplicative techniques described above.

In logarithmic pooling, the masses for each expert are weighted by an exponent where the exponents sum to one. It is not a combination to evaluate their conflicting results so there is no need to calculate the crossing cases or intersections as done in the above

examples. Equal weighting with two experts from above ([0.045, 0.255, 0.7] and [0.06, 0.24, 0.7]) uses an exponent of ½ as shown in the table below:

Table A.4: The terms used to combine expert assessments by logarithmic pooling

Expert A	Expert B	Intersection	Total
{C}: 0.045	{C}: 0.06	{C}	$0.045^{0.5} \times 0.06^{0.5} = 0.052$
{N}: 0.255	{N}: 0.24	{N}	$0.255^{0.5} \times 0.24^{0.5} = 0.248$
{C,N}: 0.7	{C,N}: 0.7	{C,N}	$0.7^{0.5} \times 0.7^{0.5} = 0.7$

The masses for each intersection can be summed as before but then need to be renormalized. Before renormalization, the summed masses for each set are:

{C}: 0.052 / (0.052 + 0.247 + 0.7) = 0.052

{N}: 0.247 / (0.052 + 0.247 + 0.7) = 0.247

{C,N}: 0.7 / (0.052 + 0.247 + 0.7) = 0.7

In this case, the normalization (0.052 + 0.247 + 0.7) is very close to one, so the values do not change much after normalization. That is not always the case.

These numbers fall between the two expert assessments but they are not a simple average. Unlike averaging, logarithmic pooling has a property called external Bayesianity. If each of the two experts were to combine their assessments with a third expert before combining with each other, you would get the same answer as if they combined their assessments at the start before combining with the third expert.

Endnotes

[1] Arvind Narayanan and Sayash Kapoor, "AI Existential Risk Probabilities Are Too Unreliable to Inform Policy," AI as Normal Technology, July 26, 2024, https://www.normaltech.ai/p/ai-existential-risk-probabilities; Ezra Karger, Josh Rosenberg, Zachary Jacobs et al., "Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament," https://forecastingresearch.org/s/XPT.pdf.

[2] Josh Rosenberg, Ezra Karger, Avital Morris et al., "Roots of Disagreement on AI Risk: Exploring the Potential and Pitfalls of Adversarial Collaboration" (Foreign Research Institute, March 2024), https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/65ef1ee52e64b52f145ebb49/1710169832137/AIcollaboration.pdf.

[3] Baobao Zhang, Noemi Dreksler, Markus Anderljung et al., "Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers," arXiv preprint arXiv:2206.04132 (2022), https://arxiv.org/abs/2206.04132.

[4] Secretary Donald Rumsfeld and General Richard Myers, "DoD News Briefing," February 12, 2002, https://web.archive.org/web/20160406235718/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636.

[5] Armen Der Kiureghian and Ove Ditlevsen, "Aleatory or Epistemic? Does It Matter?," *Structural Safety* 31, no. 2 (2009), https://www.sciencedirect.com/science/article/pii/S0167473008000556.

[6] Scott Ferson and Lev R. Ginzburg, "Different Methods Are Needed to Propagate Ignorance and Variability," *Reliability Engineering & System Safety* 54, no. 2–3 (1996), https://www.sciencedirect.com/science/article/abs/pii/S0951832096000713.

[7] Terje Aven, "The Risk Concept—Historical and Recent Development Trends," *Reliability Engineering & System Safety* 99 (March 2012), https://www.sciencedirect.com/science/article/pii/S0951832011002584; Terje Aven, "Risk Assessment and Risk Management: Review of Recent Advances on Their Foundation," *European Journal of Operational Research* 253 (August 2016), https://www.sciencedirect.com/science/article/pii/S0377221715011479.

[8] Mikita Belesni, Rusheb Shah, and Marius Hobbhahn, "Frontier Models Are Capable of In-Context Scheming," arXiv preprint arXiv:2412.04984 (2025), https://arxiv.org/pdf/2412.04984.

[9] Kenneth Payne, "AI Arms and Influence: Frontier Models Exhibit Sophisticated Reasoning in Simulated Nuclear Crises," arXiv preprint arXiv:2602.14740 (2026), https://arxiv.org/pdf/2602.14740.

[10] A.D. Moivre, "De sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus," *Philosophical transactions of the Royal Society of London* 27, 329 (1711): 213–264.

[11] Z. Zhang and C. Jiang, "Evidence-Theory-Based Structural Reliability Analysis With Epistemic Uncertainty: A Review," *Structural Multidisciplinary Optimization* 63 (2021): 2935–2953, https://doi.org/10.1007/s00158-021-02863-w.

[12] Glenn Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, 1976), https://doi.org/10.2307/j.ctv10vm1qb.

[13] Judea Pearl, "Reasoning With Belief Functions: An Analysis of Compatibility," *International Journal of Approximate Reasoning* 4, no. 5–6 (1990): 363–389, https://doi.org/10.1016/0888-613X(90)90013-R.

[14] Glenn Shafer, "Belief Functions," https://glennshafer.com/assets/downloads/rur_chapter7.pdf.

[15] Ezra Karger, Josh Rosenberg, Zachary Jacobs et al., "Forecasting Existential Risks: Evidence From a Long-Run Forecasting Tournament," https://forecastingresearch.org/s/XPT.pdf.

[16] Josh Rosenberg, Ezra Karger, Avital Morris et al., "Roots of Disagreement on AI Risk: Exploring the Potential and Pitfalls of Adversarial Collaboration" (Forecasting Research Institute, 2024), https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/65ef1ee52e64b52f145ebb49/1710169832137/AIcollaboration.pdf.

[17] Josh Rosenberg, Ezra Karger, Zach Jacobs et al., "Belief Updating in AI-risk Debates: Exploring the Limits of Adversarial Collaboration," *Risk Analysis* 45, no. 12 (2025), https://onlinelibrary.wiley.com/doi/10.1111/risa.70023.

[18] Karl Sentz and Scott Ferson, "Combination of Evidence in Dempster-Shafer Theory" (Sandia National Laboratories, April 2002), https://www.stat.berkeley.edu/~aldous/Real_World/dempster_shafer.pdf; E. Lefevre, O. Colot, and P. Vannoorenberghe, "Belief Function Combination and Conflict Management," *Information Fusion* 3, no. 2 (June 2002): 149–162, https://doi.org/10.1016/S1566-2535(02)00053-2; Didier Dubois and Henri Prade, "A Survey of belief Belief Revision and Updating Rules in Various Uncertainty Models," *International Journal of Intelligent Systems* (1994), https://doi.org/10.1002/int.4550090105.

[19] Franz Dietrich and Christian List, "Probabilistic Opinion Pooling," in *The Oxford Handbook of Probability and Philosophy* (Oxford: Oxford Academic, 2017), 519–542, https://academic.oup.com/edited-volume/43657/chapter/365891921.

[20] C. Genest, S. Weerahandi, and J. V. Zidek, "Aggregating Opinions Through Logarithmic Pooling," *Theory and Decision* 17 (1984): 61–70, https://link.springer.com/content/pdf/10.1007/BF00140056.pdf.

[21] Scott Ferson, Vladik Kreinovich, Lev Ginzburg et al., "Constructing Probability Boxes and Dempster-Shafer Structures" (Sandia National Laboratories, January 2003), https://www.cs.utep.edu/vladik/2003/sandia03.pdf.

[22] Lofti A. Zadeh, "A Simple View of the Dempster-Shafer Theory of Evidence and Its Implication for the Rule of Combination," *AI Magazine* 7, no 2, (1986), https://doi.org/10.1609/aimag.v7i2.542.

[23] Judea Pearl, "Reasoning With Belief Functions: An Analysis of Compatibility," *International Journal of Approximate Reasoning* 4, no. 5–6 (1990): 363–389, https://doi.org/10.1016/0888-613X(90)90013-R.