*Proceedings of the Royal Society, A.* 5 April 2017

http://rspa.royalsocietypublishing.org/content/473/2200/20160890

# Does a Better Model Yield a Better Argument? An Info-Gap Analysis

Yakov Ben-Haim
Yitzhak Moda'i Chair in Technology and Economics
Technion — Israel Institute of Technology
Haifa 32000 Israel
yakov@technion.ac.il

## Contents



**Abstract** Theories, models, and computations underlie reasoned argumentation in many areas. The possibility of error in these arguments, though of low probability, may be highly significant when the argument is used in predicting the probability of rare high-consequence events. This implies that the choice of a theory, model or computational method for predicting rare high-consequence events must account for the probability of error in these components. However, error may result from lack of knowledge or surprises of various sorts, and predicting the probability of error is highly uncertain. We show that the putatively best, most innovative and sophisticated argument may not actually have the lowest probability of error. Innovative arguments may entail greater uncertainty than more standard but less sophisticated methods, creating an innovation dilemma in formulating the argument. We employ info-gap decision theory to characterize and support the resolution of this problem and present several examples.

\papers\model-argum2016\tmc007.tex 2.3.2017

**Keywords** Reasoned argument, modeling, uncertainty, high-consequence events, info-gap uncertainty, robustness, satisficing.

# 1 Introduction

Three components underlie reasoned arguments of many sorts: fundamental theories, specific models derived from those theories, and computations. Rarely are these components apodictic and hence they may err. Great efforts are usually made to remove errors from our arguments, so the likelihood of error is usually very small. However, if the argument is used in predicting an extremely rare event then even tiny probability of error in the argument impugns its usefulness. When the rare event is of high consequence — meteor strike of the earth, nuclear reactor explosion, pandemic of a new disease, etc. — the implications of even tiny probability of unsound reasoning may be far-reaching.

An error in an argument is different from an error in the conclusion or implication of the argument: the conclusion may be valid even though the argument is flawed. We are concerned with error of the argument because we can't have confidence in the conclusion if we lack confidence in the argument. Warrant for the conclusion is impugned by even very small probability of unsound reasoning, and warrant increases as probability of unsoundness decreases. A new and innovative theory, or a complex and sophisticated model, or a complicated and multi-dimensional computation may be more prone to error than well-understood theories, standard models and simple computations. This has implications for one's choice of theory, model and computation in predicting rare high-consequence events.

An error of argument can result from imperfect understanding of the processes involved, or from surprising developments unknown when the argument was formulated or other obscure mechanisms. We extend earlier work by Ord, Hillerbrand and Sandberg [1], and demonstrate how info-gap decision theory can be employed in assessing the probability of erroneous argument. We show how the info-gap robustness to severe uncertainty may lead to a reversal of preference between different theories, models or computational methods used in constructing an argument. Specifically, a new and innovative theory, model or computation may be more uncertain and may have greater probability of error. Should one use the new and innovative but also more uncertain components, or the more standard and familiar ones? The analyst may thus face an "innovation dilemma" in formulating the argument, and the info-gap robustness is useful in its resolution [2].

Info-gap theory has been employed in a wide range of applications in engineering, economics, medicine, biological conservation, national security and other areas (see info-gap.com). Info-gap theory has been used in the past to evaluate the usefulness of models in engineering analysis and design [3]. This paper, however, entails a more detailed and sophisticated info-gap analysis of uncertain reasoning. Furthermore, the potential applicability of this analysis is illustrated by a range of illustrative examples.

Section 2 presents a partial survey of sources of unsound argument, and section 3 explains the importance of examining the probability of soundness and provides a framework for assessing the uncertainty of an argument. Section 4 formulates the info-gap robustness function for evaluating soundness of an argument and section 5 discusses its basic properties. Section 6 discusses several examples.

## 2 Sources of unsound argument

Unsound reasoning, even in objective, quantitative science-based argument, may arise in various ways as illustrated by the following discussion, which does not purport to be comprehensive.

Analysts may simply overlook a known and relevant factor. For example, the Castle Bravo nuclear test in 1954 yielded about 15 megatons equivalent of TNT rather than the anticipated 4 to 8 megatons. For comparison, the Hiroshima nuclear bomb released about 15 kilotons. This very significant difference in outcome — excess energy equivalent to between 470 and 730 Hiroshima bombs — resulted from failing to account for a nuclear reaction involving lithium-7 of which the analysts were aware [1].

Analysts may also err due to unknown phenomena. For instance, Parker and Risbey [4] argue that complex systems may exhibit surprising behaviors — unforeseen by modelers — because they result from processes that are not included in the analysis perhaps because they are "completely unforeseen" (p.5).

Analysts must sometimes depend on expert judgment that might err. For instance, Gregory and Lichtenstein [5] discuss a 1986 US Department of Energy study of nuclear reactor safety that "relies in part on expert assessments of the probability that something might go wrong." One scenario of future possibilities was "an 'everything else' category, called 'Unexpected Features'". They refer to earlier research, in the cognitive psychology of human judgment, indicating a strong tendency that "expert assessments of the probabilities of unlisted, unnamed events are seriously underestimated." (p.220).

Even modest errors in functional relationships may lead to significant errors in model outputs or predictions. For example, Adamson and Morozov [6] emphasize that models in biology can be very sensitive to small variations in functional relationships, and that modelers often focus only on uncertainty in the values of parameters.

Errors of reasoning can arise from quite plausible preconceptions. While the preconception may be valid, when it is combined with error in understanding the underlying process, the preconception can have far reaching implications. For example, Gunn *et al.* [7] consider the interpretation of large samples that seem overwhelmingly to support one conclusion rather than another. They argue that "even with very low systemic failure rates [e.g. error of understanding], high confidence is surprisingly difficult to achieve" (p.1). In an illustrative example they consider an ancient pot coming from either Britain or Rome, where only British producers introduced a particular chemical element in their pots. Repeated random tests indicating the presence of this element would strengthen the argument for British origin, converging on certainty as the size of the sample increases, even if most pots are Roman. Now suppose that, unknown to the analysts, a small number of Roman producers used the element thought to come only from Britain. Unanimity of test results no longer supports only British origin; the pot could be Roman. Recognition of the possibility of a very small error undermines the high confidence in the putative interpretation.

Reasoned argumentation is not a monolithic category of thought, but rather depends on deliberate or implicit choice of axioms, rules of inference, and propensities for one or another type of reason or evidence. For instance, Nisbett and Masuda [8] discuss differences between East Asian and Western modes of reasoning. They demonstrate different degrees of emphasis on logical deduction as distinct from contextual interpretation of meaning, and

on individuality of objects rather than relationship between entities. Without adopting too strong a social determinism, the results of Nisbett and Masuda remind us that scientific modelers function within a social and cultural milieu that may tend to introduce unchallenged assumptions and preferences in their choices of theories and models. Similarly, Cox [9] asserts that "experts . . . [may] reach an unwarranted consensus that replaces acknowledgment of uncertainties and information gaps with groupthink" (p.1607). Unchallenged foundations are a potential source of error in argumentation.

# 3 The importance of sound argument

Reasoned argument, often based on scientific understanding, is rightfully used to support important decisions. However, as Barrett and Danenberg [10] show in discussing States' response to the dangers of climate change, national decision makers may be unwilling to take decisive action if the threshold for catastrophe is too uncertain. It is the uncertain validity of the scientific argument that inhibits acting on the scientific prediction. Similarly, Fischhoff and Davis [11] argue, in discussing the importance of communicating scientific uncertainty, that "[d]ecision makers who place too much confidence in science can face unexpected problems, not realizing how wary they should have been. Decision makers who place too little confidence in science can miss opportunities" (p.13664). Once again, it is under- or over-estimation of uncertainty in scientific argumentation that impacts decisions. In this section we introduce a framework for assessing uncertainty of an argument, based on Ord, Hillerbrand and Sandberg [1] (subsequently OHS), that is extended in the following sections by employing info-gap decision theory.

OHS explain the very important "three-fold distinction between an argument's theory, its model and its calculations." (p.192) This is a refinement of the distinction that is widely discussed in engineering communities between 'validation' ("Have you calculated the correct equations?" which is the first 2 elements of the triplet) and 'verification' ("Have you calculated the equations correctly?" which is the 3rd element).

OHS define $X$ as a specific event, such as a specific catastrophe or some other high-consequence occurrence (p.192). We are particularly interested in events $X$ that, if they are very rare, we can either ignore $X$ or accept current contingency plans for managing $X$ should it occur. OHS consider arguments one might bring to calculate the probability that $X$ will occur (or perhaps has occurred in the past). They define $A$ as the event that the argument (used in evaluating the probability of $X$) is sound. The soundness of the argument depends on the theories that are used, how those theories are transformed into specific models, and how those models are implemented numerically. OHS then define the following entities and treat them as random variables. $T$ is the event that the involved theories are adequate, $M$ is the event that the derived models are adequate, and $C$ is the event that the calculations are correct. Using the definition of conditional probability they conclude that the probability that the argument is sound can be expressed as:

$$P(A) = P(T)P(M|T)P(C|M,T) \tag{1}$$

OHS then assume that $C$ is independent of $M$ and $T$, which is a useful simplification, though may sometimes not be justified. With this assumption eq.(1) becomes:

$$P(A) = P(T)P(M|T)P(C) \tag{2}$$

Recall that $X$ is a very rare high-consequence event whose probability $P(X)$ we wish to evaluate, and that $P(A)$ is the probability of soundness of the argument used in evaluating $P(X)$. The probability that the argument is **not** sound is $P(\neg A) = 1 - P(A)$. As we will explain shortly, we would tend to lack confidence in the argument evaluating $P(X)$ if $P(\neg A)$ is too large, especially if $P(\neg A)$ is large compared with $P(X)$. Conversely, we would give warrant to the evaluation of $P(X)$ if $P(\neg A)$ is small (or zero as one often assumes).

Why do we want $P(\neg A)$, the probability of unsoundness of the argument evaluating $P(X)$, to be on the order of $P(X)$ or less? Recall that $X$ is very pernicious but, if we accept the assertion '$P(X)$ is very tiny', then we are confident in ignoring $X$ or in accepting the current contingency management of $X$. But if the assertion '$P(X)$ is very tiny' may err with a probability far greater than $P(X)$, then the confidence derived from the smallness of $P(X)$ is overwhelmed by our lack of confidence in the argument evaluating $P(X)$.

For instance, the WASH-1400 reactor safety study reported the probability per year of an individual fatality arising from a collection of 100 nuclear power plants to be $P(X) = 2 \times 10^{-10}$ per year [12]. We might rest assured with such a small likelihood, but if the likelihood were 10 or 100 times greater we might not accept the risk. Now further suppose hypothetically, regarding the WASH-1400 calculation, that the probability that the argument is not sound is $P(\neg A) = 10^{-7}$. This is 500 times larger than the estimated $P(X)$. We would not accept the risk of $X$ if its probability were this large, so we should not accept this large a probability that our argument is wrong. To accept $10^{-7}$ as the probability that our (presumably best) argument is wrong, is similar (though not identical) to accepting $10^{-7}$ as the probability that '$P(X) = 2 \times 10^{-10}$' is wrong.[1] But if '$P(X) = 2 \times 10^{-10}$' is wrong, then action may be needed.

Thus our estimate of $P(A)$ and its complement, $P(\neg A)$, are crucial in establishing the warrant for an estimate of a very rare high-consequence event. In particular, it is important to establish that $P(\neg A)$ is very small. A very small value of $P(\neg A)$, or equivalently, a very large value of $P(A)$, constitutes strong warrant for the estimated value of $P(X)$.

# 4 Formulating the info-gap robustness of $P(A)$

We want to use $P(A)$ in eq.(2) in assessing the warrant for an estimate of the probability of a rare high-consequence event $X$, and about which there may be severe uncertainties (large info-gaps). Specifically, we will give warrant to the reported value of $P(X)$ (the theory-based estimate of the probability of $X$), if and only if $1 - P(A)$ is not large compared to $P(X)$. We will use the info-gap concept of robustness against uncertainty [13] to evaluate the warrant of the report. We begin with a simple illustration of the basic concepts.

For notational convenience, let us denote the three probabilities on the righthand side of eq.(2) as $p_1$, $p_2$ and $p_3$ (from left to right). Let us suppose that we know estimated values of the three probabilities, which we denote $\widetilde{p}_i$, $i = 1, 2, 3$. Let us furthermore suppose that we know error estimates for these three numbers, denoted $s_i$, $i = 1, 2, 3$. That is, our source for the $\widetilde{p}_i$'s has said something like: "These are my best estimates, but they could err by plus or minus $s_i$, *or more*." For instance, the estimator may have more confidence in evaluating

[1] If no other argument could justify '$P(X) = 2 \times 10^{-10}$', then $P(\neg A)$ equals the probability that '$P(X) = 2 \times 10^{-10}$' is wrong.

the adequacy of the theory (for which several Nobel Prizes have been awarded) than in evaluating the adequacy of the derived models, which are motivated in part by plausible assumptions and the desire for tractability. However, realistic bounds on the errors are unavailable. Given this information, an info-gap model for uncertainty[2] in the true values of the $p_i$'s is the following unbounded family of nested sets of $p$ vectors:

$$\mathcal{U}(h) = \left\{ p : \; p_i \in [0,\, 1],\; \left| \frac{p_i - \widetilde{p}_i}{s_i} \right| \leq h,\; i = 1,\, 2,\, 3 \right\}, \quad h \geq 0 \tag{3}$$

$\mathcal{U}(h)$ is the set of all mathematically legitimate probabilities $p_1$, $p_2$ and $p_3$ which deviate fractionally from their estimates, $\widetilde{p}_1$, $\widetilde{p}_2$ and $\widetilde{p}_3$, by no more than $h$. The value of $h$ is unknown so there is no known worst case, and the sets $\mathcal{U}(h)$ become more inclusive as $h$ grows. The family of sets is unbounded in the space of probability vectors $p$.

Like all info-gap models of uncertainty, this very simple "fractional error" model displays two properties:

$$\text{Contraction:} \qquad \mathcal{U}(0) = \{\widetilde{p}\} \tag{4}$$

$$\text{Nesting:} \qquad h < h' \implies \mathcal{U}(h) \subseteq \mathcal{U}(h') \tag{5}$$

'Contraction' means that, in the absence of uncertainty ($h = 0$), the estimate, $\widetilde{p}$, is correct. 'Nesting' means that the range of possibilities increases as $h$ increases. Together, these properties endow $h$ with its meaning as an 'horizon of uncertainty'.

Our estimated warrant — the probability that the argument is sound, $P(A|\widetilde{p})$ — is based on eq.(2), which depends on the estimated probabilities, $\widetilde{p}$. Let us suppose that this estimated warrant is large enough to make us confident in the tiny reported value, $P(X)$. Consequently we will not take any action, either because $X$ can be ignored or because we are confident in the contingency plans if $X$ occurs. Furthermore, suppose that any value of $P(A)$ greater than some critical value $P_{\rm c}$ would similarly motivate us to be confident and to take no further action to mitigate $X$.

The problem is that we really don't believe the estimated probabilities, $\widetilde{p}$, so we don't have much confidence in the estimated probability of soundness, $P(A|\widetilde{p})$. We would really like to know the true value $P(A|p)$, but we don't know the true probabilities, $p$. The "*robustness question*" that we face is: how wrong could our estimated probabilities be, and the probability of soundness would still induce us to refrain from action? If $\widetilde{p}$ could err greatly and still result in large probability of soundness, then we are robust to error in $\widetilde{p}$ and we are justified in taking no action. On the other hand, if even small errors in the estimated probabilities could undermine the warrant for $P(X)$, then perhaps we should take action to mitigate $X$.

In info-gap parlance, our "*performance requirement*" is that $P(A|p)$ be no less than the critical value:

$$P(A|p) \geq P_{\rm c} \tag{6}$$

Recall that we are considering the case that the estimated warrant, $P(A|\widetilde{p})$, satisfies this requirement. An analogous development exists for the complementary case that the estimated warrant does not satisfy eq.(6).

[2]There are many types of info-gap models of uncertainty, and we're using a very simple one here.

Finally, the *robustness* to uncertain probabilities is the greatest horizon of uncertainty, $h$, up to which the warrant $P(A|p)$ satisfies the performance requirement, eq.(6), for all realizations of the probabilities $p$ in the info-gap model $\mathcal{U}(h)$ in eq.(3):

$$\widehat{h}(P_{\rm c}) = \max\left\{h:\ \left(\min_{p\in\mathcal{U}(h)} P(A|p)\right) \geq P_{\rm c}\right\} \tag{7}$$

# 5 Interpreting the info-gap robustness of $P(A)$

A simple numerical example will provide understanding of what we learn from the info-gap robustness analysis. The mathematical details of evaluating the robustness in this example are developed in appendix A.

Let our estimated probabilities and their errors be:

$$\widetilde{p} = (0.9999,\, 0.99995,\, 0.9997), \quad s = (0.0002,\, 0.0004,\, 0.0015) \tag{8}$$

Thus $P(T)$ is estimated to equal $0.9999$, with an estimated error of $\pm 0.0002$ or more, $P(M|T)$ is estimated as 0.99995 $\pm$0.0004 or more, and $P(C)$ is estimated as 0.9997 $\pm$0.0015 or more.

The robustness curve is shown in fig. 1 for the values in eq.(8). This curve illustrates two properties of all info-gap robustness curves: zeroing and trade off.

*Zeroing:* The robustness vanishes at the estimated value of the warrant $P(A|\widetilde{p}) = \widetilde{p}_1\widetilde{p}_2\widetilde{p}_3 = 0.9996$. In other words, the estimated warrant is not reliable because its robustness to uncertainty in the probabilities is zero. Because $P(A|\widetilde{p})$ is our best estimate, one might be tempted to base one's actions on its value. However, this is unjustified due to the zeroing property and the uncertainty in the underlying probabilities. Treating $P(A|\widetilde{p})$ as the basis for action (or inaction) is wishful thinking.

*Trade off:* The negative slope of the robustness curve in fig. 1 expresses the trade off between the required performance, $P_{\rm c}$, and the robustness, $\widehat{h}(P_{\rm c})$, for achieving that performance. Good performance (large $P_{\rm c}$, strong warrant) entails poor robustness against uncertainty (low $\widehat{h}(P_{\rm c})$). This is intuitive and sometimes called the "pessimist's theorem". Combined with the zeroing property it means that we must ascend the robustness curve from its zero value, reducing our requirement in order to obtain robustness against uncertainty. The quantitative trade off, expressed by the slope of the robustness curve, allows us to assess what values of warrant, $P_{\rm c}$, can be reliably asserted because they have large robustness against uncertainty.

Now let us ascribe a 'story' to the estimated probabilities in eq.(8). Let's suppose that they refer to the newest and best theory $T$, and a sophisticated high-dimensional model implementation $M$, and a high-power parallel processing computation, $C$, needed to evaluate the model. We will call this the "*innovative*" argument and we have ascribed fairly large probabilities $\widetilde{p}$. However, because it is new, innovative, and unfamiliar, the error estimates $s$ are also rather large.

Now consider a different choice of the estimated probabilities and their errors, which represent a more conventional theory $T'$, a more standard model $M'$, and less sophisticated computation $C'$. We will refer to this argument as the "*state of the art*" (SotA). It's not the newest and best argument, but it is widely used and well understood. This argument has

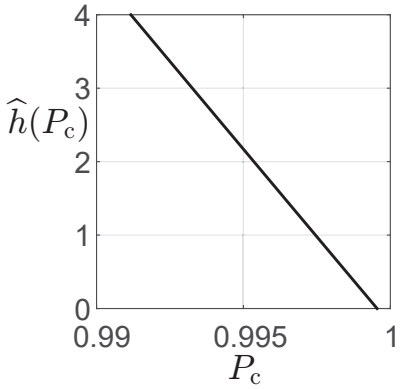


Figure 1: Robustness curve with values in eq.(8).

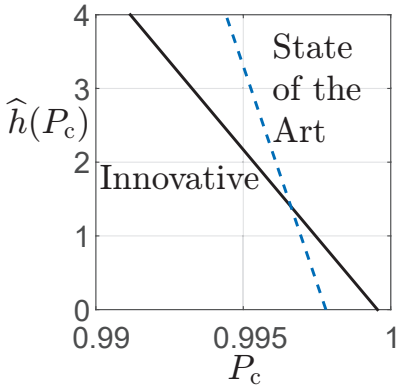


Figure 2: Robustness curves with values in eqs.(8) and (9).

lower component probabilities of being sound than the innovative argument. However, because it is the current state of the art, its error estimates are also relatively lower:

$$\widetilde{p}' = (0.9993,\, 0.9995,\, 0.9990), \quad s' = (0.0001,\, 0.00025,\, 0.0005) \tag{9}$$

Fig. 2 shows the robustness curves for both the innovative and the state of the art arguments. We see the zeroing and trade off properties of both curves.

The estimated warrant (horizontal intercept of the robustness curve) is lower for the SotA than for the innovative argument. However, because of the zeroing property, we know that this putative preference for the innovative argument does not imply that we should prefer it over the SotA argument. Specifically, the robustness of these estimated probabilities of soundness is zero in both cases, so these estimations are not a good basis for choosing between the arguments.

We note that the trade off between robustness and performance is steeper for the SotA than the for innovative case. This implies that the "*cost of robustness*" is lower for the SotA: a smaller decrement in $P_c$ is required with the SotA in order to "purchase" a given increment in robustness. The lower cost of robustness for SotA, and its lower estimated warrant, cause the robustness curves to intersect one another.

We see that the innovative argument is more robust than the SotA for $P_c$ values between $0.9965$ (at which the curves cross) and $0.9996$ (where the robustness of innovative argument becomes zero). The innovative argument will be robust-preferred over the SotA argument if the required $P_c$ is in this range. The SotA is more robust, and hence preferred, for lower values of warrant. This intersection between the robustness curves thus entails the potential for a reversal of preference between the two arguments, depending on the analyst's required warrant, $P_c$.

In summary, this example illustrates the info-gap critique of best-model optimization. Because of the zeroing property, we do not determine our preference between the arguments (innovative or SotA) based on the best estimates of their probabilities of soundness. Rather, we satisfice the warrant at $P_c$ and choose the argument that is more robust for the required warrant. This is called *robust-satisficing*. Furthermore, we use the info-gap robustness function to assess the trade off between immunity to uncertainty (robustness, $\widehat{h}(P_c)$) and performance (critical probability of soundness, $P_c$).

# 6 Model refinement and soundness of argument

## 6.1 Introduction

We now explore a slightly more realistic and interesting application of the ideas of warrant and robustness discussed above.

Consider a system that can be conceptualized as made up of discrete parts. This may be a system with functionally defined sub-systems, like a power plant with steam generators, turbines, etc. Or it may be a continuous elasto-plastic mechanical system that can be represented by a finite-element model with a discrete mesh. Or it may be a dynamic system whose behavior is discretized in time. An argument has been used to predict the probability of a rare high-consequence event, $X$, and we are concerned with the event, $A$, that this argument is sound. In particular, we are interested in evaluating the probability that the argument is sound, $P(A)$.

We will suppose that the theoretical characterization of the system is thoroughly known, so $P(T)$, the probability that the theory is adequate, is confidently known to equal one.

However, many alternative discrete models are possible. The functional units in the power plant can be resolved at many different levels, from the nut-and-bolt level all the way up to large sub-units. Likewise, the finite element representation may use rough or coarse mesh, or a temporal discretization may use small or large time steps. Let $N$ denote the degree of refinement of the model, which one could think of as the number of sub-units in the model. Our task is to choose the degree of refinement. Let $M_N$ denote the event that a model with $N$ sub-units is adequate. One might expect that the probability, $P(M_N|T)$, that an $N$-sub-unit model is adequate given the theory, will increase as the degree of refinement increases. However, the actual functional dependence of $P(M_N|T)$ on $N$ is highly uncertain and need not be strictly monotonic.

Let $C_N$ denote the event that the computation (underlying the estimation of the probability of the rare high-consequence event) is correct. Generally speaking we might expect that more interactions, and more different types of interaction, become involved, as the degree of refinement, $N$, increases. This suggests that the computational difficulty will increase as the number of model sub-units increases. Conversely, we might expect that the probability that the computation is correct, $P(C_N)$, decreases as $N$ increases. Once again, however, the actual functional dependence of $P(C_N)$ on $N$ is highly uncertain and need not be strictly monotonic.

In summary, we find that the degree of refinement of the model, as expressed by the choice of the number of sub-units, $N$, presents us with a dilemma. A finely resolved model (large $N$) will tend to have higher probability of adequacy, but likewise will tend to have lower probability of computational correctness. If we knew the dependence of $P(M_N|T)$ and $P(C_N)$ on $N$ then we could find the degree of refinement with maximal probability of soundness, $P(A)$. However, these functions are highly uncertain. Consequently, we will use info-gap theory to satisfice the probability of soundness and maximize the immunity to ignorance. This will be the basis for choosing the degree of refinement of the model used in evaluating $P(X)$.

## 6.2 First Example

We begin with a simple preliminary example. We suppose that we know estimates of the functional dependence of $P(M_N|T)$ and $P(C_N)$ on $N$, but our ignorance of the correct functional forms is represented with info-gap models of uncertainty. We evaluate the robustness to this uncertainty, and use this to prioritize the degree of model refinement, $N$.

The estimate of $P(M_N|T)$ is the following monotonically increasing function of $N$:

$$\widetilde{P}(M_N|T) = 1 - \mathrm{e}^{-\phi N} \tag{10}$$

where $\phi$ is a known positive constant. This is a stylized relation whose veracity is highly uncertain. We will use a fractional-error info-gap model to represent the uncertainty in its functional form:

$$\mathcal{U}_M(h) = \left\{ P(M_N|T) : \; P(M_N|T) \in [0,\, 1],\; \left| P(M_N|T) - \widetilde{P}(M_N|T) \right| \le h \widetilde{P}(M_N|T) \right\}, \quad h \ge 0 \tag{11}$$

This info-gap model, $\mathcal{U}_M(h)$ for $h \ge 0$, is an unbounded family of nested sets of functions $P(M_N|T)$, displaying the properties of contraction and nesting, eqs.(4) and (5). These properties imply that there is no known worse case. Note that membership of a function $P(M_N|T)$ in the set $\mathcal{U}_M(h)$ does not require the function to be monotonic.

Our estimate of $P(C_N)$ is based on the idea that the probability of computational success is very large for small and moderate refinement, $N$, falling rapidly at large refinement. Specifically:

$$\widetilde{P}(C_N) = \frac{1}{1 + \exp[\lambda(N - N_0)]} \tag{12}$$

where $\lambda$ and $N_0$ are known positive constants. The info-gap model for uncertainty in this estimate is:

$$\mathcal{U}_C(h) = \left\{ P(C_N) : \; P(C_N) \in [0,\, 1],\; \left| P(C_N) - \widetilde{P}(C_N) \right| \le h \widetilde{P}(C_N) \right\}, \quad h \ge 0 \tag{13}$$

This info-gap model also displays contraction and nesting and implies no knowledge of a worst case. Note that membership of a function $P(C_N)$ in the set $\mathcal{U}_C(h)$ does not require the function to be monotonic.

The overall info-gap model is the Cartesian product of $\mathcal{U}_M$ and $\mathcal{U}_C$, by which we mean the set of all pairs of functions, one from each set:

$$\mathcal{U}(h) = \mathcal{U}_M(h) \times \mathcal{U}_C(h) \tag{14}$$

The probability that the argument is sound, eq.(2), is:

$$P(A_N) = P(M_N|T) P(C_N) \tag{15}$$

where we have assumed that $P(T) = 1$.

The performance requirement is that the probability of soundness of the argument at refinement $N$, $P(A_N)$, must be no less than the critical value $P_\mathrm{c}$:

$$P(A_N) \ge P_\mathrm{c} \tag{16}$$

The robustness of a model with refinement $N$ and performance requirement $P_{\rm c}$ is the greatest horizon of uncertainty, $h$, up to which all realizations of the uncertain functions in the info-gap model $\mathcal{U}(h)$ satisfy the performance requirement:

$$\widehat{h}(N, P_{\rm c}) = \max\left\{h:\ \left(\min_{(P(M_N|T),P(C_N))\in\mathcal{U}(h)} P(A_N)\right) \geq P_{\rm c}\right\} \tag{17}$$

The following explicit expression for the robustness is derived in appendix B:

$$\widehat{h}(N, P_{\rm c}) = 1 - \sqrt{\frac{P_{\rm c}}{\widetilde{P}(A_N)}} \tag{18}$$

or zero if this is negative, which occurs if the performance requirement, $P_{\rm c}$, exceeds the estimated probability of soundness, $\widetilde{P}(A_N)$, which equals $\widetilde{P}(M_N|T)\widetilde{P}(C_N)$.

We illustrate these results with the following numerical example, in which $\phi = 0.3$, $\lambda = 0.2$ and $N_0 = 50$ in eqs.(10) and (12).

Fig. 3 shows plots of the estimated probabilities vs. the number of sub-units. The estimated probability that the model is adequate, $\widetilde{P}(M_N|T)$ in eq.(10), increases as the number of sub-units increases. In contrast, estimated probability that the calculation is correct, $\widetilde{P}(C_N)$ in eq.(12), decreases as the number of sub-units increases. Consequently, the estimated probability that the argument is sound, $\widetilde{P}(A_N)$, rises from low $N$ and falls off at large $N$. The greatest probability of soundness equals 0.995 and occurs with 21 sub-units.

Of course, the estimated probabilities in fig. 3 are highly uncertain, so one cannot confidently conclude that a refinement of $N = 21$ has a probability of soundness of 0.995, or that this refinement is better than others. We address this by considering the robustness to uncertainty as a function of model refinement.

Robustness curves for $N = 10, 21,$ and $38$ sub-units are shown in fig. 4. We see that the $N = 21$ refinement is more robust than the other options over the range shown (and beyond). This supports the putative preference for $N = 21$ in fig. 3, but this is only the beginning of the analysis.

The zeroing property asserts that the robustness precisely equals zero when the critical probability, $P_{\rm c}$, equals the estimated probability of soundness, $\widetilde{P}(A_N)$ in fig. 3. For instance, while $N = 21$ has a putative probability of soundness of 0.995, the robustness for this value is precisely zero as seen from the horizontal intercept of the '$N = 21$' curve in fig. 4. In other words, assessing this refinement in terms of its estimated probability of soundness is unreliable.

This leads us to consider the trade off between robustness and performance as expressed by the slope of the robustness curve. For instance, suppose we are willing to require that the probability of soundness, $P(A)$, exceed the less demanding critical value $P_{\rm c} = 0.95$. The '$N = 21$' curve in fig. 4 shows that the robustness for this requirement is 0.024. Referring to the info-gap models in eqs.(11) and (13), this means the probability of argument-soundness is guaranteed to be no less than 0.95 if the estimated probability functions, $\widetilde{P}(M_N|T)$ and $\widetilde{P}(C_N)$, err by no more than plus or minus 2.4%. Similarly, requiring probability of soundness no less than 0.90 entails a robustness of 0.048, meaning that 4.8% error in the estimated probability functions does not jeopardize a 0.90 probability of soundness.

The question “How much robustness is enough?” is a difficult value judgment depending on experience, contextual understanding, costs, consequences, and so on. Roughly speaking, if our understanding is that the estimated probabilities could err “by one percent or more”, then robustness of 0.048 is perhaps “large” and one could confidently predict a probability of no less than 0.90 that the argument is sound. On the other hand, if our understanding is that the estimated probabilities could err “by tens of percent or more”, then only very low probability of soundness can be confidently predicted.

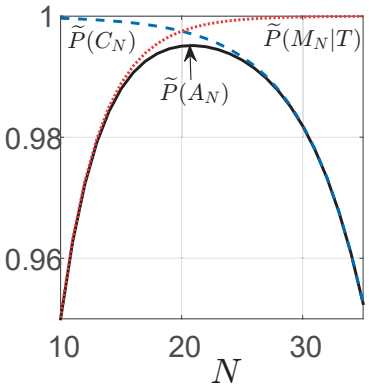


Figure 3: Estimated probabilities.

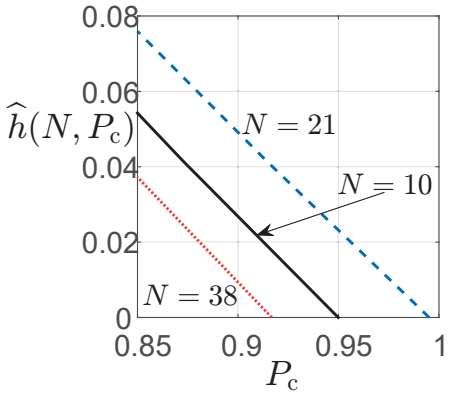


Figure 4: Robustness curves, eq.(18), for 3 values of $N$.

## 6.3 Second Example

The presumption in section 6.2 was that, for any degree of refinement $N$, a single model type was employed with estimated probability of adequacy $\widetilde{P}(M_N|T)$ in eq.(10), and that a single computational method was used with estimated probability of correctness $\widetilde{P}(C_N)$ in eq.(12). We also recognized that $\widetilde{P}(M_N|T)$ and $\widetilde{P}(C_N)$ are uncertain, as expressed by the info-gap models $\mathcal{U}_M(h)$ and $\mathcal{U}_C(h)$ in eqs.(11) and (13).

We now extend that discussion to consider a dispute about the model type and computational method. Models can come in many forms: regressions, neural nets, heuristic artificial intelligence, foundational basic science, etc. Likewise, computational methods can be diverse, including genetic algorithms, Monte Carlo, spatial-temporal discretization of scientific relations, etc. Considerable uncertainty may surround each option, as represented by info-gap models of uncertainty.

We consider two pairs of models and computations, with associated estimated probability functions and info-gap models of uncertainty. $\widetilde{P}(M_N^j|T)$ is the best estimate of the probability of type-$j$ model validity, while $\widetilde{P}(C_N^j)$ is the best estimate of the probability of correctness of the associated computation, where $j = 1$ or $2$. Likewise, the corresponding info-gap models for these entities are denoted $\mathcal{U}_M^j(h)$ and $\mathcal{U}_C^j(h)$, specified as follows:

$$\mathcal{U}_M^j(h) = \left\{ P(M_N^j|T) : \; P(M_N^j|T) \in [0,\, 1], \; \left| P(M_N^j|T) - \widetilde{P}(M_N^j|T) \right| \le h s_M^j \right\}, \quad h \ge 0 \tag{19}$$

$$\mathcal{U}_C^j(h) = \left\{ P(C_N^j) : \; P(C_N^j) \in [0,\, 1], \; \left| P(C_N^j) - \widetilde{P}(C_N^j) \right| \le h s_C^j \right\}, \quad h \ge 0 \tag{20}$$

We have introduced known positive “uncertainty weights” $s_M^j$ and $s_C^j$ as a modest generalization of eqs.(11) and (13). The overall info-gap model for type $j$ is the Cartesian product of these info-gap models and is denoted $\mathcal{U}^j(h)$, in analogy to eq.(14).

We will suppose that the estimated probabilities for type 1 are larger than for type 2:

$$\widetilde{P}(M_N^1|T) > \widetilde{P}(M_N^2|T) \quad \text{and} \quad \widetilde{P}(C_N^1) > \widetilde{P}(C_N^2) \tag{21}$$

for a relevant range of refinements, $N$. These relations reflect the fact that the type 1 model and method of computation are new, innovative and putatively better than type 2. This implies a putative preference for type 1. Nonetheless, the type 2 model and method of computation are the accepted and widely used state of the art for which experience is greater and uncertainty less than for type 1, so the uncertainty weights for type 2 are lower:

$$s_M^1 > s_M^2 \quad \text{and} \quad s_C^1 > s_C^2 \tag{22}$$

Eqs.(21) and (22) present a dilemma in the choice between the two models and computations: type 1 is putatively better but more uncertain, and hence may be worse than type 2. This is an example of an innovation dilemma, mentioned earlier [2].

The probability of argument-soundness, using the type-$j$ model and computation, is $P(A_N^j)$, the analog of eq.(15), and the performance requirement is the analog of eq.(16). The robustness of type $j$ with refinement of degree $N$ is, in analogy to eq.(17):

$$\widehat{h}(j, N, P_{\rm c}) = \max\left\{h : \left(\min_{(P(M_N^j|T), P(C_N^j)) \in \mathcal{U}^j(h)} P(A_N^j)\right) \geq P_{\rm c}\right\} \tag{23}$$

Let $\mu_j(h)$ denote the inner minimum in eq.(23), which is the inverse of the type-$j$ robustness function. That is, a plot of $h$ vs. $\mu_j(h)$ is the same as a plot of $\widehat{h}(j, N, P_{\rm c})$ vs. $P_{\rm c}$. The inverse of the robustness function is readily shown to be:

$$\mu_j(h) = \left(\widetilde{P}(M_N^j|T) - s_M^j h\right)^+ \left(\widetilde{P}(C_N^j) - s_C^j h\right)^+ \tag{24}$$

where the truncation function, $x^+$, is defined in appendix A.

The estimated probabilities of the state of the type $j$ system, in analogy to eqs.(10) and (12), are:

$$\widetilde{P}(M_N^j|T) = 1 - {\rm e}^{-\phi_j N} \tag{25}$$

$$\widetilde{P}(C_N^j) = \frac{1}{1 + \exp[\lambda_j(N - N_0)]} \tag{26}$$

where relations (21) result, for $N < N_0$, from:

$$\phi_1 > \phi_2 \quad \text{and} \quad \lambda_1 > \lambda_2 \tag{27}$$

Figs. 5–9 show results for the following parameter values: Innovative: $\phi_1 = 0.3$, $\lambda_1 = 0.4$, $s_M^1 = 0.18$, $s_C^1 = 0.21$. State of the art: $\phi_2 = 0.2$, $\lambda_2 = 0.2$, $s_M^2 = 0.05$, $s_C^2 = 0.05$.

Figs. 5 and 6 shows the estimated probabilities for the innovative and State of the Art (SotA) model and computational method, respectively. The estimated probabilities are higher for the innovative than for the SotA case, which is why the innovation is attractive. We also note that the estimated probability of soundness is maximal for a different refinement in the innovative case ($N = 28$) than in the SotA ($N = 25$).

Robustness curves for 3 degrees of refinement are shown in fig. 7 for the innovative case. We see that $N = 28$ is the robust-dominant refinement for the innovative model.

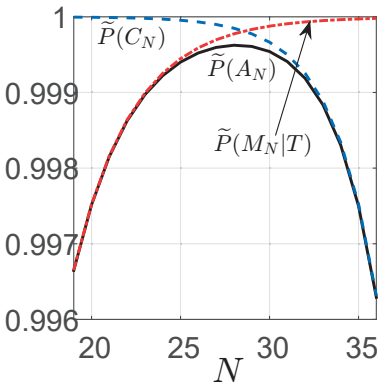


Figure 5: Estimated probabilities, system type 1 (innovative).

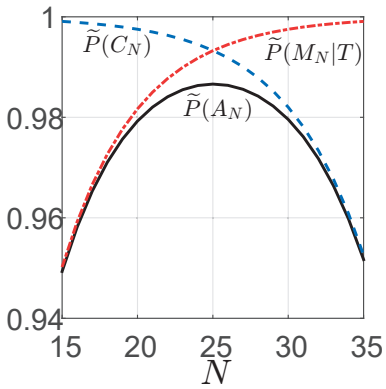


Figure 6: Estimated probabilities, system type 2 (SotA).

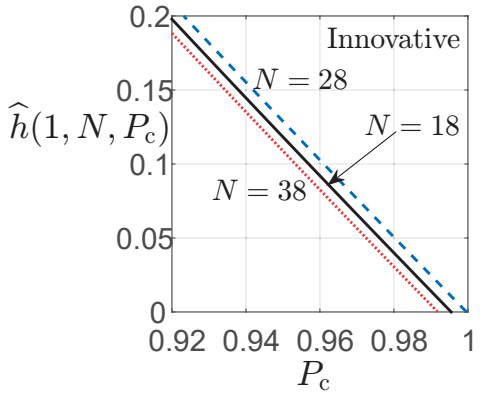


Figure 7: Robustness curves, system type 1 (innovative), from eq.(24).

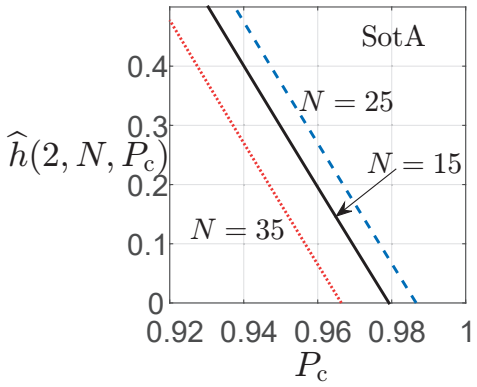


Figure 8: Robustness curves, system type 2 (SotA), from eq.(24).

However, the zeroing property shows that the robustness of achieving the estimated probability of soundness, 0.9996, is precisely zero. The trade off between robustness and performance shows, for instance, a robustness of 0.051 for achieving soundness-probability of 0.980. This means that the probability of soundness will be no less than 0.980 if the estimated probabilities err no more than plus or minus 0.051 times the corresponding uncertainty weight, $s_M^1$ or $s_C^1$. That may or may not be adequate robustness depending on various considerations discussed briefly above.

Fig. 8 shows robustness curves for 3 degrees of refinement of the SotA case. The estimated probabilities are lower than in the innovative case so the horizontal intercepts are further to the left than in fig. 7. However, the cost of robustness—expressed by the slope of the robustness curve—is lower in the SotA case: a given increment in robustness is obtained with a lower reduction in $P_c$ in the SotA case.

This is seen most clearly in fig. 9, which shows robustness curves for the best innovative and SotA options. The robustness curves cross one another, indicating that the innovative option is robust-preferred at large $P_c$, while the SotA option is robust-preferred at lower $P_c$. The intersection occurs at critical probability $P_c = 0.982$ and robustness of $\widehat{h} = 0.045$. Thus, for example, if a 0.97 probability of soundness is acceptable, then the SotA is more robust and hence preferred, while requiring 0.99 probability of soundness implies a robust preference for the innovative option. In both cases, however, one needs to decide if the

robustness is sufficiently large.

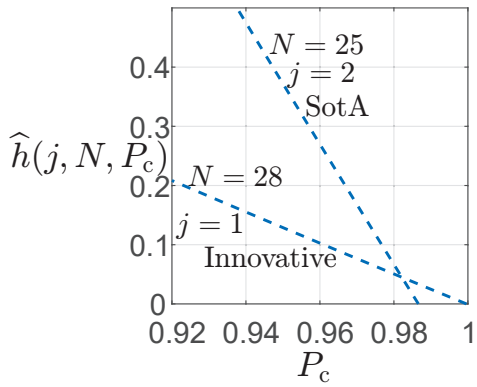


Figure 9: Robustness curves, system types 1 and 2 (innovative and SotA), from eq.(24).

**In conclusion,** we have argued that the putatively best model, theory or computational method may not be the best choice for reasoned prediction of rare high-consequence events. The possibility of error in these predictions, though of low probability, may be highly significant when the argument is used in predicting the probability of rare high-consequence events. Consequently, the choice of a theory, model or computational method for predicting rare high-consequence events must account for the probability of error in these components. However, error may result from lack of knowledge or surprises of various sorts, and predicting the probability of error is highly uncertain. We show that the putatively best, most innovative and sophisticated argument may not actually have the lowest reliably-assessed probability of error. Innovative arguments may entail greater uncertainty than more standard but less sophisticated methods, creating an innovation dilemma in formulating the argument. We illustrated the method of info-gap robust-satisficing to characterize and support the resolution of this problem.

**Ethics statement.** This work did not involve any active collection of human data.

**Data accessibility statement.** This work does not have any experimental data.

**Competing interests statement.** The author has no competing interests.

**Funding.** This research was not funded.

**Acknowledgements.** The author is indebted to valuable discussion with Toby Ord and Anders Sandberg. This paper was written while the author was at the University of Oxford as a Distinguished Visiting Scholar, an Oxford-Martin Visiting Fellow and a Leverhulme Visiting Professor. The author is grateful for the support of these foundations.

# 7 References


1. Toby Ord, Rafaela Hillerbrand and Anders Sandberg, 2010, Probing the improbable: methodological challenges for risks with low probabilities and high stakes, *Journal of Risk Research,* Vol. 13, No. 2, pp.191–205.

2. Yakov Ben-Haim, Craig D. Osteen and L. Joe Moffitt, 2013, Policy dilemma of innovation: An info-gap approach, *Ecological Economics,* 85: 130–138.

3. Yakov Ben-Haim, Scott Cogan and Laëtitia Sanseigne, 1998, Usability of mathematical models in mechanical decision processes, *Mechanical Systems and Signal Processing,* 12: 121–134.

4. Parker W.S. and Risbey J.S., 2015, False precision, surprise and improved uncertainty assessment. *Phil. Trans. R. Soc. A,* 373: 20140453. http://dx.doi.org/10.1098/rsta.2014.045

5. Robin Gregory and Sarah Lichtenstein, 1987, A review of the high-level nuclear waste repository siting analysis, *Risk Analysis,* 7(2): 219–223.

6. Adamson M.W. and Morozov A. Yu., 2013, When can we trust our model predictions? Unearthing structural sensitivity in biological systems. *Proc R Soc A,* 469: 20120500. http://dx.doi.org/10.1098/rspa.2012.0500

7. Gunn L.J., Chapeau-Blondeau F., McDonnell M.D., Davis B.R., Allison A. and Abbott D., 2016, Too good to be true: when overwhelming evidence fails to convince. *Proc. R. Soc. A,* 472: 20150748. http://dx.doi.org/10.1098/rspa.2015.0748

8. Richard E. Nisbett and Takahiko Masuda, 2003, Culture and point of view, *Proc. Natl. Acad. Sci.,* 100(19): 11163–11170

9. Louis Anthony (Tony) Cox, Jr., 2012, Confronting Deep Uncertainties in Risk Analysis, *Risk Analysis,* 32(10): 1607–1629.

10. Scott Barrett and Astrid Dannenberg, 2012, Climate negotiations under scientific uncertainty, *Proc. Natl. Acad. Sci,* 109(43): 17372–17376.

11. Baruch Fischhoff and Alex L. Davis, 2014, Communicating scientific uncertainty, *Proc. Natl. Acad. Sci,* 111, suppl. 4: 13664–13671.

12. US. Nuclear Regulatory Commission, 1975, Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants, WASH-1400 (NUREG75/014), table 1-1, p.3.

13. Yakov Ben-Haim, 2006, *Info-Gap Decision Theory: Decisions Under Severe Uncertainty,* 2nd edition, Academic Press, London.

# A Evaluating the info-gap robustness in section 4

In this section we explain how the robustness function in eq.(7) is evaluated.

Let $\mu(h)$ denote the inner minimum in the definition of the robustness function, $\widehat{h}(P_{\rm c})$, in eq.(7). Some consideration will show that $\mu(h)$ is the inverse function of $\widehat{h}(P_{\rm c})$. That is, a plot of $h$ vs. $\mu(h)$ is identical to a plot of $\widehat{h}(P_{\rm c})$ vs. $P_{\rm c}$. Thus it is sufficient if we determine the function $\mu(h)$. The robustness curves in figs. 1 and 2 were obtained by plotting $h$ vertically vs. $\mu(h)$ horizontally.

For subsequent notational convenience let us define the following truncation function. $x^+$ equals $x$ if $x \in [0,\, 1]$, it equals 0 is $x \leq 0$ and it equals 1 otherwise.

From eq.(2) and the definition of the probabilities $p_i$ we know that:

$$P(A|p) = p_1 p_2 p_3 \tag{28}$$

Consideration of the info-gap model in eq.(3) shows that $\mu(h)$ is obtained, at horizon of uncertainty $h$, by using the smallest allowed values of the $p_i$'s. Thus:

$$\mu(h) = \prod_{i=1}^{3} (\widetilde{p}_i - s_i h)^+ \tag{29}$$

This uniquely specifies the robustness function $\widehat{h}(P_{\rm c})$.

# B Evaluating the info-gap robustness in section 6.2

In this section we explain how the robustness function in eq.(17) is evaluated.

Let $\mu(h)$ denote the inner minimum in eq.(17), the definition of the robustness. Note that $\mu(h)$ is the inverse of $\widehat{h}(P_{\rm c})$. That is, a plot of $h$ vs. $\mu(h)$ is identical to a plot of $\widehat{h}(P_{\rm c})$ vs. $P_{\rm c}$.

As before, we define the following truncation function. $x^+$ equals $x$ if $x \in [0,\, 1]$, it equals 0 is $x \leq 0$ and it equals 1 otherwise. Using the info-gap models of eqs.(11), (13) and (14) we find that $\mu(h)$ occurs by choosing $P(M_N|T)$ and $P(C_N)$ as small as possible at horizon of uncertainty $h$:

$$P(M_N|T) = (1-h)^+ \widetilde{P}(M_N|T), \quad P(C_N) = (1-h)^+ \widetilde{P}(C_N) \tag{30}$$

Thus, recalling the assumption that the theory is correct, so that $P(T) = 1$, we find from eqs.(2) and (30):

$$\mu(h) = \Big[(1-h)^+\Big]^2 \underbrace{\widetilde{P}(M_N|T)\widetilde{P}(C_N)}_{\widetilde{P}(A_N)} \tag{31}$$

$\widetilde{P}(A_N)$ is the estimated probability of soundness of the argument. The robustness is the greatest value of $h$ at which $\mu(h)$ is not less than performance requirement, $P_{\rm c}$. Because $\mu(h)$ is monotonically decreasing, the robustness is obtained by solving the following relation for $h$:

$$\mu(h) = P_{\rm c} \tag{32}$$

Conbining eqs.(31) and (32) yields eq.(18).

## Figure Captions

Fig. 1. Robustness curve with values in eq.(8).
Fig. 2. Robustness curves with values in eqs.(8) and (9).
Fig. 3. Estimated probabilities.
Fig. 4. Robustness curves, eq.(18), for 3 values of $N$.
Fig. 5. Estimated probabilities, system type 1 (innovative).
Fig. 6. Estimated probabilities, system type 2 (SotA).
Fig. 7. Robustness curves, system type 1 (innovative), from eq.(24).
Fig. 8. Robustness curves, system type 2 (SotA), from eq.(24).
Fig. 9. Robustness curves, system types 1 and 2 (innovative and SotA), from eq.(24).