Journal of AppliedMath 2025, 3(4), 2992.
https://doi.org/10.59400/jam2992

Article

Model validation based on value-of-information theory

George Hazelrigg

Department of Mechanical Engineering, George Mason University, Fairfax, VA 22030, USA; ghazelri@gmu.edu

CITATION

Hazelrigg G. Model validation based on value-of-information theory. Journal of AppliedMath. 2025; 3(4): 2992.
https://doi.org/10.59400/jam2992

ARTICLE INFO

Received: 24 March 2025
Revised: 22 April 2025
Accepted: 27 April 2025
Available online: 2 July 2025

Abstract: The modeling and simulation community has devoted considerable attention to the question of model validity as a condition for the use of a model in an engineering decision-making process. Their work has focused on the concept of "accuracy", loosely defined as the difference between a model-computed result and a real-world result. The objective of this paper is to introduce an alternative approach, based on classical decision theory, that focuses on the value of the information that a model provides to the decision-making process. This is a significant departure from the current approach to model validation, and it derives from the preference, "I want the best outcome that I can get". Use is made of an example case that results in a paradox to illustrate weaknesses in the accuracy-focused approach. Instead of advocating the use of a model based on its accuracy, this work advocates using a model if it adds value to the overall application, thus relating validation directly to system performance. The approach fills significant gaps in the current theory, notably providing a clearly defined validity metric and a mathematically rigorous rationale for the use of this metric.

Keywords: decision theory; value of information; model validation; systems modeling

1. Introduction

This paper is concerned specifically with the validation of models that are intended for use in support of an engineering design decision-making process. It is noted that use of all of the current, accuracy-focused validation methods inherently reflect a preference that greater model accuracy is always better. However, this paper shows that, since accuracy is not a fundamental system design preference, it is merely a means to satisfying the real system preference, it cannot be assured that accuracy-focused validation leads to optimal design decision making and, in fact, may lead to decisions that detract from the final overall system performance. We show that this weakness of accuracy-focused validation, together with its other deficiencies can be overcome by using a value-of-information approach based on classical decision theory. On the other hand, we recognize that accuracy may be a valid preference for models that serve purposes other than decision-making.

Engineers have long been concerned with the accuracy of their models, particularly as regards their use in decision making and policy setting. Accordingly, the concept of model validation emerged together with the field of operations research beginning in the 1940's [1]. More recently, verification and validation (V&V) have become major components of modeling and simulation. This trend has been advanced by both government funding and Congressional requirements. For example, when the decision was made in 1992 to end nuclear testing, Congress created the Stockpile Stewardship Program (SSP) and charged it with an annual assessment of the "safety, security, and reliability of the nuclear weapons stockpile in the absence of nuclear testing", to be issued by the directors of the Lawrence Livermore National Laboratory, Los Alamos National Laboratory and Sandia National Laboratories (SNL) [2]. This led to significant efforts to advance the science of V&V, leading to several publications of which the following references are a sample [3–13]. The current approach to model validation was framed by these efforts and remains largely unchanged.

Despite this activity, the definitions of the terms *verification* and *validation* remain somewhat under debate. In 1968, Fishman and Kiviat [14] noted, "*validation tests whether a simulation model reasonably approximates a real system*." They go on to say, "*V&V insure that a simulation model is properly designed; only after a model has been verified and validated can an experimenter justifiably use a model to probe system behavior*." In 2012, the National Academy of Sciences [2] offered, "*The process of validation involves comparisons between QOIs* [quantities of interest] *computed by the model and corresponding true, physical QOIs inferred from physical observations or experiments. The intended use of the model determines how close the model's QOIs must be to the true QOIs in order for the model to serve its purpose; that is, the intended use determines the requirements on model accuracy.*" Thus, it is inferred that, should there be a significant difference between the model results and the real world, the model should not be used. However, there is no specific guidance offered as to what might be a significant difference. More recently, MITRE [15] defines validation as "*The process of determining the degree to which a* [simulation] *model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model.*"

Despite the remaining debate, the latter two definitions are generally accepted by the V&V community [16]. However, they fall short on two points. First, they are based on the concept of *accuracy* without a clear definition of or fundamental rationale for this measure, and second, they offer no guidance on what level of accuracy must be obtained in order that one use a particular model for a given purpose. Drawing on classical decision theory and the value of information [17–19], this paper addresses these two weaknesses and links model validation directly to a theory of systems engineering [20].

The underlying concepts upon which this work is based are well-developed [21].[1] We do not add to them here. Rather, the contribution of this work is to place these concepts to use in making the decision whether to use a particular model in support of a particular system design task, that is, to "*validate*" a model for a particular use. We begin, Section 2, by reviewing the gambler's paradox, which illustrates key weaknesses in the current accuracy-focused model validation approach. We then discuss the resolution of this paradox, Section 3, which points to additional deficiencies in the current approach. In Section 4, we introduce the concept of information in the context of a decision-making process and suggest a quantitative measure of information. Section 5 makes note of the conditions relating to probability theory that the proposed approach is dependent upon. In Section 6, we examine model validation relying on the information provided by a model and point to the limitations of using information alone to determine model validity. Section 7 completes the theory by consideration of the risk-adjusted value of information obtained by the use of a model. This section illustrates the theory using a very simple example in order that we can focus the discussion on the process of value-based model validation without the distraction that would result from the use of a complex engineering model[2] [22–29]. Despite the simplicity of the example problem, it should be noted that this is an "engineering" example and that the only difference between this example and a complex engineering example is that the model used here is easily understood, as the example illustrates all of the computations needed for more complex cases. Section 8 discusses the application of the theory to more complex engineering cases, and Section 9 presents conclusions, noting that value-focused model validation addresses all of the weaknesses of accuracy-focused validation and it can be shown to be consistent with the preference that, "We want the best outcome that we can get." One will see that, as the value-of-information theory is developed, it becomes important to modify the definition of model validity, ending with a definition that clearly links model validation to system performance.

2. The gambler's paradox

The following paradox illustrates weaknesses in the theory of accuracy-focused model validation. A gambler, about to place a series of bets on a coin-flip game, is presented with the opportunity to purchase a model that predicts the outcome of a coin flip.[3] Being skeptical, she insists on testing the model before her purchase. A very rigorous validation yields the result that the model is always wrong, that is, each time the model predicts heads, the coin lands tails, and each time the model predicts tails, the coin lands heads. The gambler must now address two questions. First, is the model valid and, second, should she purchase and use the model? According to the above definitions of "valid", she must conclude that the model is not only not valid, it could not be more in error. Hence, by current practice, she would be discouraged from purchasing and using the model. On the other hand, if she held the belief that her validation results are an accurate assessment of the future performance of the model, then she could bet against the model and expect to win on every gamble. It doesn't get better than this. So, logic tells us that she obviously should purchase and use the model.

One might be inclined to argue that, in this case, there exists an inverse model, interchanging heads and tails, that is perfectly correct, and thus there is no paradox. But suppose, instead of being perfectly incorrect, the model was wrong 80% of the time? Let us assume that heads and tails comprise the complete outcome set, that is, they are the only outcomes of the coin flip with non-zero probability, and further that the gambler believes that the coin is fair. We shall take "fair" to mean that she believes that, as the number of flips becomes large, the coin will land heads half of the time and tails half of the time with no historical correlation. Given these conditions, it may still be the case that the model provides information of value to the gambler. That is, she can still improve her odds of winning by betting against the model. Alternatively, suppose the model were correct 80% of the time? Then she could bet with the model and improve her odds of winning. Indeed, the only case in which the model adds no value at all is the case when the model is wrong 50% of the time, that is when the error rate of the model precisely reflects the gambler's belief in the probability of heads versus tails. It is only in this case that the model provides no additional information beyond the gambler's belief.

To say that the model is not "valid" as defined above is, ergo, inconsistent with the notion that a model that is not valid should not be used. How might this paradox be resolved, and what lessons might we learn in the process?

3. Resolution of the gambler's paradox

To begin, we should acknowledge that a key purpose of model validation is to provide confidence that a model can be used for guidance in a decision-making situation. *Predictive models support decision making by adding value*[4] *to the decision*. But models themselves do not make decisions, they simply present a computed "result" displayed on a computer screen or printout. Decisions are made typically by people (or other such intelligent agents) and, in order for a model result to be used in a decision-making process, it must go from its display into the head of the decision maker. Thus, there can be and usually is a discrepancy between the model result and the "information" in the head of the decision maker. That discrepancy is often associated with an assessment of the degree to which the model is an accurate representation of reality. Indeed, in order for the decision maker to make optimal use of the model result, in passing from display into the head of the decision maker, it would be modified by the a priori beliefs of the decision maker in accord with Bayes' theorem[5] [30–32].

Bayes' theorem is a rule of Kolmogorov probabilities that enables one to "update" a prior belief based on recently acquired data and the decision maker's belief about the accuracy and applicability of the data. It is stated as follows:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{1}$$

This is read, the probability of occurrence of A given that B is true is the probability of B given that A is true times the probability of occurrence of A divided by the probability of the occurrence of B. We can think of Bayes' theorem as a consistency condition on the decision maker's beliefs regarding A and B. Having conducted many tests of the coin-flip model, and believing that the coin is fair, it is consistent in this case for the gambler to believe that a model result of heads infers that tails is the better bet and that a model result of tails infers that heads is the better bet. It is shown in an example below that Bayes' theorem enables a decision maker to make optimal use of both model and validation results obtained from other sources.

This points to the next inconsistency of accuracy-focused model validation, which seeks to validate the degree to which the model produces a result that is consistent with physical reality whereas, from the decision-making point of view, it would be more appropriate to validate the decision maker's a posteriori beliefs given the model result. But this leads us to conclude that any such validation is that of a particular decision maker, that is, it is unique to that specific decision maker and does not transport to others.[6] This presents a third inconsistency regarding accuracy-focused validation, which seeks to dissociate the validation process from the decision maker. The question facing us now is, what is a logically consistent approach to determining whether a model should be used at all and, if so, what degree of "trust" should be placed in the model?

This brings us to the topic of *information*.

4. Information

In classical decision theory, *information* relates to our ability to predict the future. It is the basis for rational decision making. Engineering education focuses on teaching an ability to solve problems. But problems are quite different from decisions, and the mathematics of the two are not the same. Problems are solved, resulting in solutions (answers). Solutions are obtained typically seeking consistency with laws of nature as we know them, mathematics (logical constructs), and data or boundary conditions, and we judge solutions to be right or wrong. Problem solutions are frequently deterministic. Decisions, on the other hand, are made, not solved. In order to make a rational decision, one must have alternatives from which to choose, beliefs on the possible outcomes of each alternative, and a preference over the outcomes [33,34][7]. Decisions always result in outcomes, which are judged good or bad, and all real decisions involve uncertainty and risk. The mathematics of decision making is called decision theory, and the most widely accepted and axiomatically based normative theory is that of von Neumann and Morgenstern [35–37] often referred to as von Neumann-Morgenstern (vNM) utility theory. Decision theory seeks to optimize decisions in the context of available alternatives, beliefs (probabilities) on the plausible outcomes of these alternatives and the decision maker's preference over the outcomes of each alternative. Dating from the early 1940s, it is important to recognize that vNM utility theory builds on Kolmogorov probability theory [38]. It does not allow the use of any alternative theory of uncertainty, nor would there seem to be valid alternative decision rules based on alternative theories of uncertainty.

We shall now derive a measure of information, referred to as the *state-of-information* with Kolmogorov probability as its basis. Consider a typical decision such as that shown in **Figure 1** [21]. In this figure, the square represents a decision, and the large dots represent uncertain events.

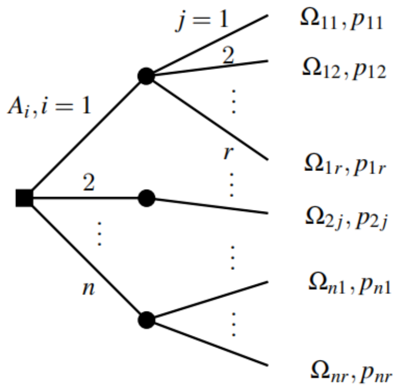

Figure 1. A graph or "decision tree" for a typical decision.

The available alternatives are denoted A_i. Each alternative, A_i, enables a range, r, of possible outcomes, Ω_{ij}, each outcome occurring with probability p_{ij}. Associated with each outcome is a real scalar "value" measure, v_{ij}, such that, given the preference

$$\Omega_{ij} \succ \Omega_{IJ} \tag{2}$$

then

$$v_{ij} > v_{IJ} \tag{3}$$

Further, given the indifference

$$\Omega_{ij} \sim \Omega_{IJ} \tag{4}$$

then

$$v_{ij} = v_{IJ} \tag{5}$$

The value measures, v_{ij}, are ordinal and must satisfy the condition that if $v_{ij} \geq v_{kl}$ and $v_{kl} \geq v_{mn}$, then $v_{ij} \geq v_{mn}$ must be true[8] [21,39]. Now define the variable

$$\delta_{ij}^{IJ} = \begin{cases} 0 & \text{if} \quad v_{ij} < v_{IJ} \\ 1 & \text{if} \quad v_{ij} \geq v_{IJ} \end{cases} \tag{6}$$

Thus, δ_{ij}^{IJ} is 0 if outcome IJ is preferred to outcome ij, otherwise it is 1. With this definition of δ_{ij}^{IJ}, the probability that alternative A_i will result in an outcome that is preferred or indifferent to alternative A_I is given by

$$p_i^I = \sum_j p_{ij} \sum_J \delta_{ij}^{IJ} p_{IJ} \tag{7}$$

In applying Equation (7), it is important to realize that the p_{ij} are conditional probabilities (Hazelrigg [21], p. 185). Let the preferred alternative be A_ι. We now determine the state of information as the probability that the following statement is true:

> *Alternative A_ι will produce an outcome that is preferred or indifferent to all other available alternatives.*

This probability is

$$H_\iota = \prod_{I \neq \iota} p_\iota^I \tag{8}$$

This definition of the state of information is such that, if $H_\iota = 1$, the decision maker feels certain that alternative A_ι will provide an outcome that is preferred or indifferent to the outcomes that would result from any of the other available alternatives. The value of H_ι decreases as the state of information is degraded. If a choice from among n alternatives is such that all alternatives have the same probability of achieving the most preferred outcome, one could expect $H_\iota = 1/n$. Thus, if the a posteriori value of $H_\iota = 1/n$, we would not expect that the extant information would be of value.

Clearly, given this definition of information, the closer H_ι is to unity, the more confident that the decision maker can be that his or her preferred choice will yield the best available outcome. Yet, while we refer to the condition $H_\iota = 1$ as a state of *perfect information*, note that a state of perfect information does not infer a state of zero uncertainty[9] [40]. Thus, we see that it is possible that we can be certain of making the best choice even in the context of model uncertainty on the outcomes of the available alternatives.

5. Notes on probability

At this point, it is again important to emphasize that the above mathematics is dependent on the use of Kolmogorov and only Kolmogorov probability. No attempt should be made to extend value-based validation to any alternative concept of uncertainty. vNM utility theory and the value-of-information theory are built upon Kolmogorov probability, and we know of no other normative decision theory that is consistent with the Dutch Book Argument, which is widely accepted by the mathematics community as a necessary condition for a self-consistent theory of uncertainty [30–32, 38, 41–48]. Unfortunately, the reader will note that the literature deviates frequently from accepted practice [7, 8, 11, 49, 50]. Results from such works are not consistent with decision theory and cannot be relied upon to provide valid conclusions. Kolmogorov theory treats probability as a belief [19], that is, as a mathematical concept devoid of any physical properties. Ergo, there is no such thing as a physically "correct" probability estimate. This again reinforces the notion that model validity is a belief held by a specific decision maker in regard to a specific decision.

6. Information validation

The H measure derived above provides an alternative approach to model validation. Although this measure does not ensure that a model presents an accurate representation of the real world, it provides a self-consistent measure of the degree of trust that the decision maker holds in basing her decisions on a state-of-information that includes a model result. In fact, we can use the H measure to map out confidence levels relating to a specific model's use in a specific decision situation.

Another point that should be made is that, in the case of a decision, we don't need to know how well each alternative will perform (that might be of interest, but it is not the point in making the decision), rather the question is, which alternative will yield the most preferred outcome, and what risks[10] are we willing to accept to get that outcome? What we need the model to help us with is *resolving the differences* between the outcomes and the risks associated with the alternatives.[11]

Given this discussion, we might define model validity in a way that is more consistent with the above notions:

> *A model is valid if, when used in a specific decision making situation with a given set of available alternatives and the decision maker's beliefs and preferences, the decision maker is certain that his/her preferred choice is the choice that indeed yields the outcome that is most preferred from among the outcomes that would have been obtained from any of the other available alternatives.*

More simply, but less accurately, we define a model to be valid if its use assures the decision maker that the preferred choice will result in the best possible outcome. This definition places the notion of model validity directly into the context of a decision.

However, it is not very useful in the common case where there is uncertainty regarding which alternative is the preferred choice. Thus, we now return to the concept of *information* in a more rigorous context, which will lead ultimately to a more refined definition of validity.

The measure H_ι gives the probability, as perceived by the decision maker, that alternative ι will provide the most preferred outcome from among the alternatives available. However, we may not assume that alternative ι is the preferred alternative merely because $H_\iota \geq H_i$ for all $i \neq \iota$. That would require comparison of expected utilities. It may be the case that alternative ι has a remote possibility of a disastrous consequence that the decision maker wants to be sure of avoiding, and hence will not choose this alternative. This case is eloquently illustrated by the St. Petersburg paradox posited by Nicolas Bernoulli in 1713. The paradox involves a game in which a coin is flipped n times until it lands heads. The player then receives a prize of $\$2^n$. Thus, the expected payoff of the game is

$$\mathrm{E}\{\text{Prize}\} = \sum_{n=1}^{\infty} \$2^n \frac{1}{2^n} = \$1 + \$1 + \$1 + ... = \$\infty \tag{9}$$

The question now is, how much would you pay to play this game? The minimum prize is \$2. So, surely a person would pay that amount. But the expected value is infinite, and it is equally clear that no one would pay an infinite sum to play the game. In fact, if they did, they would lose with probability 1. Indeed, most players would pay about \$4 to play the game. A solution to this paradox, proposed by Daniel Bernoulli [51], was published in the Commentaries of the Imperial Academy of Science of Saint Petersburg in 1738.

The St. Petersburg paradox makes it clear that we must take into account the risk preference of the decision maker. We do this by means of a von Neumann-Morgenstern lottery [21]. A vNM lottery is a gamble in which the decision maker is presented with a choice between the receipt of a certain outcome, A, and a lottery, B, whose outcome is dependent on an uncertain event such as a coin flip. The choice of B results in either of two outcomes, B_1 or B_2, where $B_1 \succ A \succ B_2$. The decision maker is then asked for the probability of obtaining outcome B_1, $p(B_1)$, such that she is indifferent between A and the lottery B. Given the outcomes A, B_1 and B_2 and the probability $p(B_1)$, and noting that utility is indifferent to an affine transformation, namely that two of the utility values can be chosen arbitrarily with the condition that $u(B_1) > u(A) > u(B_2)$, we can compute three points on a value curve consistent with the decision maker's risk preference. If the outcomes are all monetary, this computation yields the decision maker's risk-adjusted preference for money.

$$u(A) = p(B_1)u(B_1) + [1 - p(B_1)]u(B_2) \tag{10}$$

where, for example, $u(A)$ is referred to as the "utility" of A as shown in **Figure 2**. Lotteries such as this can be posed iteratively to better define the decision maker's utility curve.

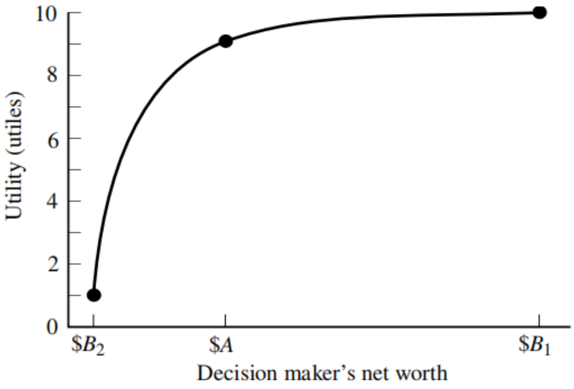

Figure 2. Utility versus money for the above example.

Given the decision maker's utility curve as defined above, risky alternatives can be ranked according to their expected utilities, namely if alternative A is preferred to alternative B and using the abbreviated nomenclature $\mathrm{E}\{u_A\} = \mathrm{E}\{u(A)\}$, then $\mathrm{E}\{u_A\} > \mathrm{E}\{u_B\}$ where, for example, $\mathrm{E}\{u_B\}$ is given by the right hand side of Equation (10) and, since A is deterministic, $\mathrm{E}\{u_A\} = u_A$. Thus, the expected utilities, $\mathrm{E}\{u_i\}$, enable a deterministic mapping of alternatives onto $\mathbb{R}^1$ such that alternatives are ranked in accordance with the decision maker's preference regarding both plausible outcomes and risk.[12]

Particularly in the design of artifacts or systems that have continuous design variables **x**, we would expect an optimal design (the maximum expected utility design) that has many neighboring designs that are only slightly less preferred than the optimal design.[13] We could also expect that, as a result of uncertainty, it might be the case that one of these neighboring designs, if chosen, would actually yield an outcome that is preferred to the outcome obtainable from the maximum expected utility design. Also, there might be other designs, not near the optimum design, that are only slightly less preferred than the optimal design. This follows from the notion of global and local maxima. Again, it might be impossible for a model to resolve with complete confidence which region of the design is truly optimal. So, our concern with models includes both their ability to resolve the optimal design in a particular neighborhood, and to resolve which of different neighborhoods contains the globally optimal design.

Suppose that we consider a choice between two design points, the maximum expected utility or preferred design, $\mathbf{x}_O$, and an alternative design, $\mathbf{x}_C$. For this simple choice between only two alternatives, if $H_{O/C} = 1$, the decision maker is confident that point $\mathbf{x}_O$ will yield a more preferred outcome than point $\mathbf{x}_C$. Thus, we segment the design space into two parts by defining the set M_O as the set of all points $\mathbf{x}_C$ for which pairwise comparisons yield $H_{O/C} < 1$, the shaded areas in **Figure 3**. We would expect that at least a subset of M_O will be in the neighborhood of point $\mathbf{x}_O$, and there is the possibility that there are additional subsets of M_O. We will designate the subsets of M_O as M_{Oq}, where M_{O0} refers to the subset in the neighborhood of design point $\mathbf{x}_O$ (all of the points in M_{O0} are connected contiguously with the point $\mathbf{x}_O$), which we shall call the near subset, and other subsets, $q = 1, 2, 3, \ldots$, which we shall refer to as the global subsets of M_O.

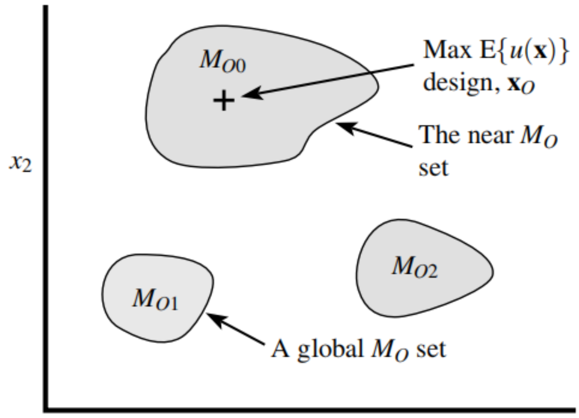

Figure 3. Illustration of the M_O set in a two-dimensional design space.

Of course, confidence-level sets could also be drawn for conditions such as $H_{O/C} \geq 0.9$. These may provide useful insights in certain circumstances.

7. The value of information

At this point, we can formulate a criterion for the determination of the usefulness of a model as a function of the value of the information that it provides to the decision [17]. Specifically, we shall define a model to be "valid" if it adds net value to the decision. *This is consistent with the preference that we want the best system (or product design) that we can get.*[14] *Accuracy-focused validation fails to assure this condition and can, in fact, lead to the use of models that degrade system performance or deny the use of models that would improve system performance.* Suppose that we face a decision in which alternative A_ι is the preferred alternative, that is, it is the alternative with the greatest expected utility, $\mathrm{E}\{u(A_\iota)\}$. Let's further suppose that the decision maker has a net worth of $\$w$ and a risk preference utility for money given by $u_{dm} = \ln(w)$.[15] The decision maker may choose simply to select alternative A_ι or, alternatively, choose to obtain additional information on the set of alternatives, which might focus only on analyzing alternative A_ι in more detail. Note that the alternatives, A_i, may include remaining at the status quo, sometimes thought of as the "do-nothing" alternative. It would be advantageous to seek additional information if doing so would increase the expected utility of the decision situation.

As a quick note here, it would not be advantageous to seek additional information if $H_{\iota/C} = 1$ for all $\mathbf{x}_C$ despite uncertainty in the outcomes as this value of H denotes complete confidence in the mind of the decision maker that the preferred choice will result in the most preferred outcome achievable and thus additional information would have no value. The case of $H_{\iota/C} = 1$ for all $\mathbf{x}_C$ occurs when the probability distribution of the outcome of alternative A_ι does not overlap with the probability distribution of the outcome of any other alternative. The comedian Eddie Izzard illustrates this case by posing the question, "Cake or death"?[16] Here, the choice is obvious despite considerable uncertainty on the outcomes of both alternatives, and the case of $H_{\iota/C} = 1$ for all $\mathbf{x}_C$ is a "cake-or-death" decision situation.

We begin by looking at the case where $H_{\iota/C} < 1$ for some $\mathbf{x}_C$ and the decision is made to gather added information by means of an analysis. Prior to the analysis, we have probability distributions on the possible outcomes of each alternative, $p_i(\Omega_i)$ with elements $p_{ij}(\Omega_{ij})$ (refer to **Figure 1** for nomenclature). Subsequent to the analysis, we have computed probability distributions, $p^{*}_{ij}(\Omega_{ij})$. Our next step is to update our prior distributions given the distributions $p^{*}_{ij}(\Omega_{ij})$. With discrete probabilities such as would be obtained by a computer simulation, we do this by invoking Bayes' theorem, Equation (1), in the form

$$p^{+}_{ij}(\Omega_{ij}) = p_{ij}[(\Omega_{ij})|p^{*}_{ij}(\Omega_{ij})] = \frac{[p^{*}_{ij}(\Omega_{ij})|p_{ij}(\Omega(_{ij})]p_{ij}(\Omega_{ij})}{\sum_{j} p_{ij}(\Omega_{ij})[p^{*}_{ij}(\Omega_{ij})|p_{ij}(\Omega_{ij})]} \tag{11}$$

where $p^{+}_{ij}(\Omega_{ij})$ are the a posteriori probabilities of the outcomes Ω_{ij}. The use of improved information then takes the form of replacing the $p_{ij} = p_{ij}(\Omega_{ij})$ with the $p^{+}_{ij}(\Omega_{ij})$ in **Figure 1**.

Equation (11) provides several insights of which we should take note. First, it is clear that we would not accept the result of the analysis in the absence of an assessment of its validity as given by $p^{+}_{ij}(\Omega_{ij})$. Second, it is not appropriate to merely accept the result of the analysis with total disregard to our a priori assessment, $p_{ij}(\Omega_{ij})$, except under the special condition that we place no confidence at all in our a priori assessment, which is to say that we have an uninformed or uniformly distributed prior over the entire range of plausible outcomes. Instead, we update our a priori assessment given the result of our analysis. Third, the Bayesian update asks that we make statements about our confidence in our a priori assessment and the result of our analysis. These statements comprise the subjective opinion of the decision maker. Therefore, Equation (11) provides a subjective a posteriori assessment of the decision maker's state-of-information, which could be interpreted loosely as a decision maker's accuracy-focused statement of validation.

A word of caution is in order here. It would not be unusual to seek improved information from a variety of sources either simultaneously or sequentially. In either case, a desired property of the method by which a posteriori probabilities are computed is that the result not be path dependent. That is, the final a posteriori probability estimate should not depend on the order in which information improvements are accounted. Bayesian probability updating has the property that it is not path dependent. This might not be the case with alternative methods. Thus, it is strongly recommended that only Bayesian updating be used.

At this point, a simple example might help. The example chosen here uses a very simple model intentionally in order to maintain a clear focus on the mathematics of the ensuing validation process. The application of this process to more complex engineering cases is merely a matter of substituting a different, more complex model in place of that used in this example. It follows that this example presents all of the mathematical operations needed to analyze complex engineering cases. Consider being offered the opportunity to play a game in which you are asked to place a bet on how many candies are in a jar filled with M&Ms, **Figure 4**.[17] You are required to pay an entry fee, $\$F$, to play the game, which is played against other players. The winner is the person whose guess is closest to the exact number of M&Ms in the jar, but that does not exceed the exact number by even one M&M. The winner receives a monetary prize, $\$P$. Though quite simple, this game is remarkably similar to engineering design decision making where one wants the most economical design possible, but where a design that is the slightest bit too economical can result in catastrophe.[18]

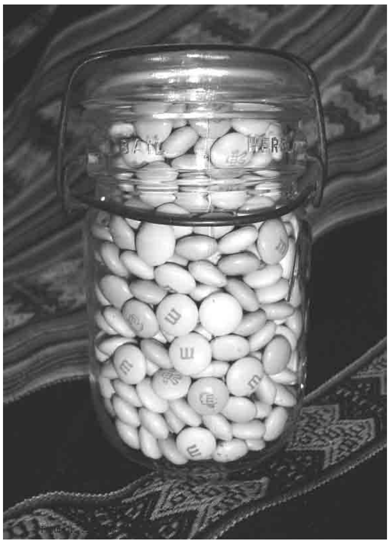

Figure 4. A jar full of M&Ms.
Note that the way in which the M&Ms pack into the jar creates considerable uncertainty as to how many M&Ms are actually in the jar.

As a first step, you could simply guess at how many candies might be in the jar. Perhaps we would characterize your guess by a most likely value (your best guess), a minimum value (you are certain that at least this many M&Ms are in the jar), and a maximum value (you are certain that not more than this number of M&Ms are in the jar), and we could assume that a triangular distribution adequately represents your beliefs. Then, we do a simulation of candies in the jar. For this, we could do a deterministic computation and believe that this number is correct to, say, plus or minus a factor of 2, **Figure 5**, or we could perform a Monte Carlo simulation to actually compute a probability distribution. In either case, a convenient model is

$$N = \frac{\eta V_j}{V_c} \tag{12}$$

where N is the estimated number of M&Ms in the jar, η is a packing factor (the volume of the candy in the jar divided by the volume of the jar), V_j is the volume of the jar and V_c is the average volume of an M&M. We then apply Equation (11) to compute the posterior distribution, **Figure 6**.

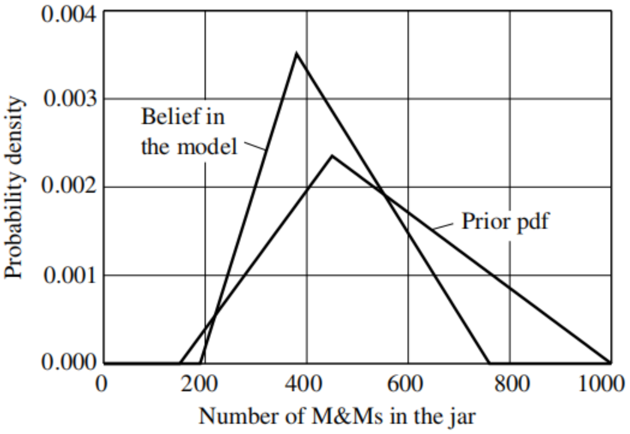

Figure 5. Probability density functions representing a prior belief on M&Ms in the jar and a belief in the model.

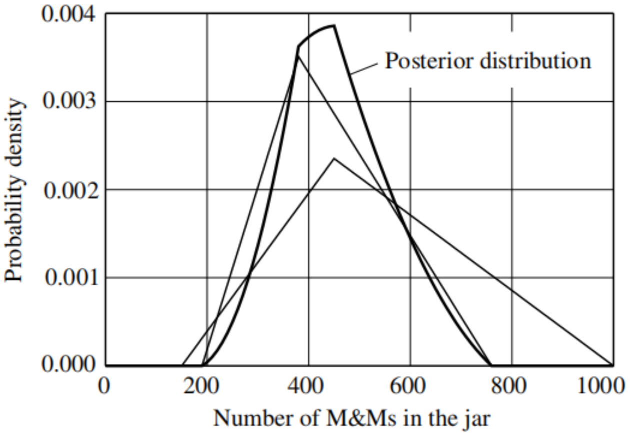

Figure 6. Posterior belief on M&Ms in the jar.

An important concept in this game is that the object of the game is not to guess exactly how many M&Ms are in the jar, the object is to win the prize. Thus, the expected utility of the game is

$$\mathrm{E}\{u_{\text{game}}\} = p_{\text{win}} u_{\text{win}} + (1 - p_{\text{win}}) u_{\text{lose}} \tag{13}$$

where, assuming that you prefer to win (you are free to want to lose), $u_{\text{win}} > u_{\text{lose}}$. It follows that the best guess maximizes the probability of winning, p_{win}. You now get to choose whether or not to play the game.[19] A flowchart for computer simulation of the M&M game is provided in the Appendix. Noting that the decision not to play maintains the deterministic status quo, the decision would be to play if your expected utility of playing is greater than your utility of the status quo, that is, if $\mathrm{E}\{u_{\text{game}}\} > u_w$. We are now in a position to assess the value of the analysis.

First, however, we can make another interesting observation. From Equation (13), we can see that, if the player wants to lose, then the optimal solution would be to maximize p_{lose}, and although the player would have rather imperfect information had she wanted to win, our model clearly enables her to make a guess that is guaranteed to lose. For example, she could choose Avogadro's number as the guess with absolute certainty of losing. Thus, we see that the value, ergo also the validity, of a model depends on the decision maker's preference as well as her beliefs.

Improving the state-of-information can create value in two ways. First, it can lead to choices that result in outcomes of higher value. For example, it can lead to design optimizations that come closer to the ideal optimal design. Second, an improved state-of-information can lead to a higher probability of a payoff, which is to say it can reduce risk. Even without increasing the expected value of the game, reduction of risk can be of monetary value to a decision maker who is risk averse, that is, who has a concave downward utility curve such as shown in **Figure 2**. An improved state-of-information benefits a player in the M&M game described above by reducing risk, that is, by increasing the probability of winning the game.

Given your initial state of information and a guess at the actions of the other players in the game, you can compute a number of M&Ms, N_O, that has the highest probability of winning, p_{win}. Notice that there are two decision points in this game, first, whether to play or not and, second, if playing, what number of M&Ms to guess. At each decision point, we take the choice with the highest expected utility, and we base the expected utility of the game on these choices being taken. We now compute the expected utility of the game with your a priori state-of-information. Recall for our example we are taking your utility to be given by $\ln(w)$ so that $u_{\text{lose}} = \ln(w-F)$ and $u_{\text{win}} = \ln(w-F+P)$. Thus, the expected utility of the game is

$$\mathrm{E}\{u_{\text{game}}\} = p_{\text{win}} \ln(w-F+P) + (1-p_{\text{win}}) \ln(w-F) \tag{14}$$

If $\mathrm{E}\{u_{\text{game}}\} > \ln(w)$, you would choose to play the game, and if $\mathrm{E}\{u_{\text{game}}\} < \ln(w)$, you would choose not to play the game. In the latter case, your utility would remain $\ln(w)$. A simulation of the game could yield an a posteriori maximum utility guess, N_O^+, that has the highest probability of winning, p_{win}^+. Then, if the decision were to play, the a posteriori expected utility of the game would be

$$\mathrm{E}^+\{u_{\text{game}}\} = p_{\text{win}}^+ \ln(w-F+P-C) + (1-p_{\text{win}}^+) \ln(w-F-C) \tag{15}$$

where C is the cost of performing the analysis. Again, if $\mathrm{E}^+\{u_{\text{game}}\} > \ln(w)$, you would choose to play the game, and the value of the information gained by the analysis would be given by

$$\$V = e^{\mathrm{E}^+\{u_{\text{game}}\}} - e^{\mathrm{E}\{u_{\text{game}}\}} \tag{16}$$

Alternatively, if a choice not to play the game is not changed by the a posteriori belief, $\mathrm{E}^+\{u_{\text{game}}\} = \mathrm{E}\{u_{\text{game}}\} = \ln(w)$, and the value of the information would be $\$V = \$0$ (that is, remaining at the status quo means that the model offers no additional value). On the other hand, were the simulation to reverse a decision to play the game, then the simulation would have positive value.

Note here that Equations (14) and (15) include the player's wealth as a parameter. Ergo, we see that this too impacts the assessment of the value of the analysis and the validity of the model.

Let's put some numbers to this example so that we can see the impact of the decision maker's risk preference. We shall assume that there are 20 players in this game, all equally naive. So, we might choose to estimate the decision maker's a priori probability of winning to be $1/20 = 0.05$. We will also assume the following: the decision maker's current wealth, w, is \$1000; the entry fee to the game, F, is \$200; the cost of an analysis, C, is \$200; and the prize, P, is \$5000. Finally, we will take the computed a posteriori probability of winning, p_{win}^{+} to be 0.3. Given the a priori estimate of winning, the utility of winning is

$$u_{\text{win}} = \ln(1000 - 200 + 5000) = 8.666 \tag{17}$$

and the utility of losing is

$$u_{\text{lose}} = \ln(1000 - 200) = 6.685 \tag{18}$$

Thus, the expected utility of playing the game is

$$\mathrm{E}\{u_{\text{game}}\} = 0.05 \times 8.631 + (1 - 0.05) \times 6.397 = 6.784 \tag{19}$$

Whereas the utility of the status quo is

$$u_{\text{win}} = \ln(1000) = 6.908 \tag{20}$$

Since the expected utility of playing is lower than the utility of the status quo, the decision maker would choose not to play. Having performed an analysis, the utility of winning is

$$u_{\text{win}} = \ln(1000 - 200 + 5000 - 200) = 8.631 \tag{21}$$

and the utility of losing is

$$u_{\text{lose}} = \ln(1000 - 200 - 200) = 6.397 \tag{22}$$

The expected utility of playing given the results of an analysis is

$$\mathrm{E}^{+}u_{\text{game}} = 0.3 \times 8.631 + (1 - 0.3) \times 6.397 = 7.067 \tag{23}$$

So, we see that the results of the analysis changed the decision maker's decision from not playing to playing the game. Thus, the economic value net of the cost of the analysis to the decision maker is

$$\$V = e^{\mathrm{E}^{+}\{u_{\text{game}}\}} - e^{\mathrm{E}\{u_{\text{game}}\}} = \$289 \tag{24}$$

The expected value of the game with improved information is $0.3 \times (\$5000 - \$200 - \$200) - 0.7 \times (\$200 + \$200) = \$1100$, but the risk-adjusted value difference between playing and the status quo is only \$289.

Suppose the decision maker has not made the decision to perform an analysis, but wishes to evaluate whether to perform an analysis or not. It would seem natural that the decision maker would want to perform an analysis if the cost of performing the analysis is less than its worth to the decision maker. In this case, we do not have the value of p^{+}_{win} as determined by the analysis, but we have an expectation on this value p^{*}_{win}. Thus we can estimate p^{*}_{win} and solve Equations (15) and (16) replacing p^{+}_{win} with p^{*}_{win} to see if the decision maker expects the analysis to have a positive value.

8. The value of information when the payoff is a function of the design

To begin, we note that the only significant differences between the candy-jar example and a real engineering example are the model that is being validated and the decision maker's belief in the accuracy of the model. The engineering model would replace Equation (12) , and the decision maker's belief in the accuracy of the model would replace the belief distribution shown in **Figure 5**. Again, this distribution may be a simple expression of the decision maker's belief, or it could be augmented with a detailed analysis, such as a Monte Carlo simulation accounting for the sources of uncertainty in the model. A complication in the case of an engineering example could involve obtaining the data for the analysis.[20] As seen by Equation (15), these data include the "entry fee" (that is, F, the cost of the project together with its uncertainty), the "prize" (P, namely the profit or payoff expected from the project together with its uncertainty), the cost of the analysis, C, for which there should be a reasonable estimate with relatively small or negligible uncertainty, and a determination of the decision maker's risk preference, which would include the status quo of the decision maker's (or the firm's) net worth, w. Typically, this would refer to a firm's willingness to accept financial risk. These parameters would normally be included in any assessment of a project, but it would be appropriate to include uncertainty estimation on them for the purpose at hand[21].

In engineering design and systems engineering, the payoff of a design decision is often a function of the design itself. This would be the equivalent of the case where the payoff in the M&M game is a function of the error between the winning guess and the true number of M&Ms in the jar, $P(N, n)$. In this case, the utility of the game depends on accuracy as well as the probability of winning, and we need to find the guess, N_O, that maximizes expected utility of the game as opposed to merely maximizing the probability of winning. In principle, this case is much the same as above, but the computational complexity is a bit greater. We must now also compute the utility of the game as a function of your guess for each simulation, and then average these utilities to get the expected utility of each guess.

In the case where the choice has been made to perform an analysis that has led to improved information, the analysis has provided either a deterministic performance outcome estimate or it has resulted in a probability distribution of possible performance outcomes. In the former case, it would be appropriate to assign a belief to the deterministic estimate in the form of a statement about its accuracy such as was illustrated in **Figure 5**. Given this distribution, we invoke Equation (11) to obtain a posterior distribution as shown in **Figure 6** and, using this distribution, we can compute the expected utility of winning the game with improved information resulting from the analysis. Again, if the improved information increases the expected utility of the game, we would want to accept the analysis as valid and make use of it, and we can also compute the risk-adjusted dollar value of the analysis as we did above, Equation (24).

This case illustrates one final consideration that is not apparent in the accuracy-focused validation process. If the other players in the game also have access to a computer model that will provide improved information, and if they use it, their use of the model can decrease the value of your model to you and, perhaps, invalidate its use. So, we conclude that the value of a model, hence its validity, can depend on the actions of others as well as on the model itself.

The final case that needs to be discussed is the case where we wish to evaluate the potential value of conducting an analysis or simulation where the payoff depends on the accuracy of the guess. In this case, we don't have the simulation that we used above to evaluate the analysis, and it would seem that the best we can do without getting into rather complex computations is to simply make some estimates of the improvement that the analysis might provide and proceed with a simple computation such as that given in the previous section.

9. Conclusions

The key contribution of the work presented in this paper is to illustrate shortcomings inherent to all forms of accuracy-focused model validation in the case that the model is to be used in support of a decision-making process, and to provide an alternative model validation approach based on the value-of-information theory, which derives from classical decision theory. It is noted that this approach to model validation resolves all of the deficiencies of accuracy-focused model validation.

Recognizing that we care about the accuracy of models because we intend to use them in support of decision-making, it would seem obvious that a theory of validation based on the mathematics of decision theory would be more appropriate than an accuracy-focused theory, particularly as the former is more comprehensive than and inclusive of Kolmogorov probability theory and vNM decision theory, and it is consistent with the Dutch Book Argument. Furthermore, the gambler's paradox points to a number of weaknesses in the current theory, beginning with the use of the concept of *accuracy* as the determining metric of validity. Accuracy is neither a well-defined concept nor is there a fundamental imperative for its use. Whereas model accuracy is not a fundamental preference in the design of a system (it is merely a means to an end), it can be shown that value is [20], and that the use of a value-focused validation will support decision making aimed at achieving optimal choices. Hazelrigg and Saari [20] also show that, to achieve maximum system performance, all decisions must be made based on the overall system preference, and this includes the choice of the metric that defines model validity, which must be a system value metric. Thus, we suggest defining a model to be "valid" if its use adds value to the process in which it is to be used. This does not exclude the use of an accuracy assessment of a model by an entity other than the decision maker. But it does mean that any such assessment must be mathematically valid and applied within the relevant context of the decision. In the example of the gambler's paradox, while the model would have value at the extremes of always right or always wrong, it also has value if it is right or wrong, say 75% of the time, but it has no value when it is right or wrong precisely 50% of the time. Accuracy-focused validation does not capture this behavior.

Next, we note that model uncertainty and its quantification are mathematical concepts totally devoid of physical reality [19, 41]. It is a concept that resides in the mind of an analyst and is not a property of any device or system. Thus, model validation is a personal thing, and a model that one decision maker would deem valid might not be valid in the mind of another decision maker even though both decision makers are considering the same application.

Additionally, we see that the value of a model depends on the decision maker's status quo position (or that of the organization) when making the decision as this determines her status quo utility. It also depends on the decision maker's preference, as this is a key determinant of the decision maker's state of information. In the case of the M&M game, changing a player's preference from wanting to win to wanting to lose moves the player from a position of relatively poor information to a position of perfect information.

Finally, the M&M game shows that the value of a model can depend on what other people are doing. In particular, in this game, if other people also do analyses, the information they get through their analyses has the potential to diminish the value of improved information to our decision maker, and this can affect the validity of a model as based on the value it imparts to the application.

To summarize, a value-of-information theory of model validation addresses the following shortcomings of accuracy-focused validation: (1) it never forces a treatment of uncertainty that is not consistent with accepted mathematical practice (for example, using non-Kolmogorov measures of uncertainty); (2) it does not offer measures of uncertainty for use in decision-making processes that have no corresponding mathematically valid decision framework; (3) it enables the use of beliefs that are those of the decision maker in the estimation of uncertainty; (4) it accounts for the preferences of the decision maker; (5) it provides a clear definition of what constitutes model validity for a particular decision; (6) it offers a mathematically clear definition of model validity; (7) it provides a fundamental basis for the use of value as a measure of validity; and (8) it provides a framework for testing the validity of difference models used for comparing alternatives in cases where excessively inaccurate models of the alternatives are highly correlated [52,53]. Such models frequently occur in engineering cases where a large source of uncertainty such as weather is the same for the different alternatives, while the result of the uncertainty has significantly different impact on each alternative.

Funding: This work has been supported by the National Science Foundation under award CMMI-1923164.

Conflict of interest: The author declares no conflict of interest.

Notes

[1] For clarity and completeness of this paper, we review the concepts presented in this reference here in the context of value-based model validation.

[2] For example, the Sandia V&V challenge problem is complex enough that it took a 10-page journal article to state the problem. The complexity of the mathematical model was such that none of the five solution responses, which focused on the model, detected that the problem was ill-posed. A sixth paper explored "techniques for decision analysis using economic benefit-cost analysis of model validation and nonmodeling alternatives," but fell short of applying utility theory to obtain a risk-adjusted value of a model.

[3] Indeed, a key purpose of engineering modeling and simulation is prediction of system performance in support of a decision.

[4] Such value may be tangible as in the form of money, or intangible as in the form of risk reduction.

[5] Bayes' theorem can be shown to be consistent with the Dutch Book Argument, which shows that a gambler may be subject to a certain loss unless his or her actions are entirely consistent with the axioms of Kolmogorov probability.

[6] While not as obvious, it should be noted that the same is true of model validation using a frequentist approach, which typically makes the inherent assumptions that the decision maker exerts no prior knowledge, is risk neutral and accepts the subjective IID (independence and identically distributed) assumption needed to interpret frequencies as probabilities.

[7] It is worth noting that Dodgson taught mathematics at Oxford, and his specialty was logic and decision theory.

[8] The mapping of outcomes into the real scalar measure, v, is possible only in the case that the decision maker is rational, that is, has transitive preferences. This is arguably the case for individuals, but not for groups. This reinforces the understanding that model validation applies only to a specific individual and does not transport.

[9] It is also interesting to note that, writing $S_\iota = -\ln H_\iota$ leads to an entropy interpretation of H that accounts for preferences. This measure of informational entropy extends Shannon's theory to the decision that will be made using the "received" information.

[10] In the context of decision theory, "risk" is defined as variability in the outcome of an action, leading to plausible outcomes that include both worse and better than a norm.

[11] Note that there are two different ways to resolve differences between outcomes using models. The first is to evaluate the alternatives modeling each separately and computing the differences based on the model results. The second is to model the difference directly. Very frequently, modeling the difference directly provides a significantly more accurate difference result than the former method.

[12] Note that the measure u is cardinal and, for a given alternative, its expected value, $\mathrm{E}\{u\}$, is deterministic and ordinal. The sole purpose of expected utility is to convert a stochastic optimization problem into a deterministic optimization problem in the context of a mathematically rigorous decision-making framework.

[13] The astute reader might note that, if the model has less than perfect accuracy, it may be impossible to determine the true best design point, however, the best expected utility design can be determined and it affords the best choice of the design point under the current state-of-information. We are reminded here that there is a difference between the best decision and the best outcome.

[14] In this context, the definition of "*best*" is left entirely up to the decision maker.

[15] A common approximation is that, for a decision that puts only a small fraction of the decision-maker's wealth at risk, we can assume a risk-neutral decision maker, and for a decision that puts a significant fraction of the decision-maker's wealth at risk, we can assume the log form of utility. But we would always choose to use the decision maker's personal risk preference if that is available.

[16] https://www.youtube.com/watch?v=unkWbEmtYXs Warning, this video contains adult language.

[17] The choice of M&Ms is predicated on the fact that these candies are uniform in size and shape, closely approximating oblate spheroids, thus making their volume convenient to model. And they taste good when the game is over.

[18] The Hyatt Regency Hotel walkway collapse, July 17, 1981, is an example.

[19] Persons interested in experimenting with this game may ask the author for a Windows compatible simulation.

[20] This complication would be equally present if one were performing an accuracy-focused model validation.

[21] It might seem odd that a decision maker's net worth would enter into the validation of a system model. However, it should be clear that we all take risks in a way that relates to the magnitude of the risk compared to our net worth. Hence, we demand greater certainty in cases where we might decide to put a higher fraction of our worth at risk and, thus, our determination of model validity does indeed depend on w.

References

1. Sargent RG, Balci O. History of verification and validation of simulation models. In: Proceedings of the 2017 Winter Simulation Conference; 3–6 December 2017; Las Vegas, NV, USA. pp. 292–307.
2. National Research Council. Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantifi-cation. National Academies Press; 2012.
3. Oberkampf WL, Deland SM, Rutherford B, et al. Estimation of total uncertainty in modeling and simulation. Sandia National Laboratories; 2000.
4. Easterling RG. Measuring the predictive capability of computational models: Principles and methods, issues and illustrations. Sandia National Laboratories; 2001.
5. Trucan TG, Easterling RG, Dowding KJ, et al. Description of the Sandia Validation Metrics Project. Sandia National Laboratories; 2001.
6. Hills RG, Trucano TG. Statistical validation of engineering and scientific models: A maximum likelihood based metric. Sandia National Laboratories; 2002.
7. Sentz K, Ferson S. Combination of evidence in Dempster-Shafer theory. Sandia National Laboratories; 2002.
8. Ferson S, Kreinovich V, Ginzburg L, et al. Constructing probability boxes and Dempster-Shafer structures. Sandia National Laboratories; 2003.
9. Hills RG, Leslie IH. Statistical validation of engineering and scientific models: Validation experiments to application. Sandia National Laboratories; 2003.
10. Oberkampf WL, Trucano TG, Hirsch C. Verification, validation, and predictive capability in computational engineering and physics. Sandia National Laboratories; 2003.
11. Ferson S, Hajagos J, Berleant D, et al. Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis. Sandia National Laboratories; 2004.
12. Helton JC, Johnson JD, Oberkampf WL, Storlie CB. A sampling-based computational strategy for the representation of epistemic uncertainty in model predictions with evidence theory. Sandia National Laboratories; 2006.
13. Oberkampf WL, Pilch M, Trucano TG. Predictive capability maturity model for computational modeling and simulation. Sandia National Laboratories; 2007.
14. Fishman GS, Kiviat PJ. The statistics of discrete-event simulation. Simulation. 1968; 10(4): 185–195.
15. MITRE. Systems engineering guide. MITRE; 2014.
16. Toptan A, Porter NW, Hales JD, et al. Construction of a code certification matrix for heat conduction with finite element code applications. Journal of Verification Validation and Uncertainty Quantification. 2020; 5(4): 041002–041017.
17. Howard RA. Information value theory. IEEE Transactions on Systems Science and Cybernetics, SSC. 1966; 2(1): 22–26.
18. Hillier FS, Lieberman GJ. Introduction to Operations Research, 5th ed. McGraw-Hill; 1990.
19. Howard RA, Abbas AE. Foundations of Decision Analysis. Pearson Publishing; 2016.
20. Hazelrigg GA, Saari DG. Toward a theory of systems engineering. Journal of Mechanical Design. 2022; 144: 011402–011410.
21. Hazelrigg GA. Fundamentals of Decision Making for Engineering Design and Systems Engineering. NEILS CORP; 2012.
22. Hu KT, Orient GE. The 2014 Sandia verification and validation challenge: Problem statement. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011001–011011.
23. Beghini LL, Hough PD. Sandia verification and validation challenge problem: A PCCM based approach to assessing prediction credibility. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011002–011012.
24. Choudhary A, Voyles IT, Roy CJ, et al. Probability bounds analysis applied to the Sandia verification and validation challenge problem. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011003–011016.
25. Li W, Chen S, Jiang Z, et al. Integrating baysian calibration, bias correction, and machine learning for the 2014 Sandia verification and validation challenge problem. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011004–011016.
26. Xi Z, Yang RJ. Reliability analysis with model uncertainty coupling with parameter and experiment uncertainties: A case study of 2014 verification and validation challenge problem. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011005–011016.
27. Mullins J, Mahadevan S. Bayesian uncertainty integration for model calibration, validation, and prediction. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011006–011016.
28. Hazelrigg GA, Klutke GA. Models, uncertainty, and the Sandia V&V challenge problem. Journal of Verification, Validation and

Uncertainty Quantification. 2020; 5: 015501–015507.

29. Paez PJ, Paez TL, Hasselman, TK. Economic analysis of model validation for a challenge problem. Journal of Verification, Validation and Uncertainty Quantification. 2016; 1: 011007–011020.
30. Hájek A. Scotching Dutch Books? Philosophical Perspectives. 2005; 19(1): 139–151.
31. Hájek A. Dutch Book Arguments. In: Anand P, Pattanaik P, Puppe C (ediotors). The Oxford Handbook of Rational and Social Choice. Oxford University Press; 2009
32. Caves CM. Probabilities as betting odds and the Dutch Book. Available online: http://info.phys.unm.edu/ caves/reports/dutchbook.pdf (accessed on 27 May 2025).
33. Carroll L. Alice’s Adventures in Wonderland. United Kingdom; 1865.
34. Hazelrigg GA. The Cheshire Cat on engineering design. Quality and Reliability Engineering International. 2009; 25: 759–769.
35. von Neumann J, Morgenstern O. The Theory of Games and Economic Behavior, 3rd ed. Princeton University Press; 1953.
36. Savage LJ. The Foundations of Statistics. Dover Publications, Inc.; 1972.
37. Luce RD, Raiffa H. Games and Decisions. John Wiley & Sons; 1957.
38. Bernardo JM, Smith AFM. Bayesian Theory. Wiley; 2000.
39. Arrow KJ. Social Choice and Individual Values, 2nd ed. Wiley; 1963.
40. Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1984; 27: 379–423, 623–656.
41. Lindley DV. Understanding Uncertainty. John Wiley & Sons; 2014.
42. Lindley DV. Scoring rules and the inevitability of probability. International Statistical Review. 1982; 50: 1–26.
43. Lindley DV. The philosophy of statistics. The Statistician. 2000; 49: 293–337.
44. Ramsey FP. Truth and Probability. In: The Foundations of Mathematics and other Logical Essays, Ch. VII. Harcourt, Brace and Company; 1931. pp. 156–198.
45. de Finetti B. Theory of Probability. Wiley; 1974. Volume 1.
46. de Finetti B. 1975. Theory of Probability. Wiley; 1975; Volume 2.
47. de Finetti B. Theory of Probability. John Wiley & Sons, Ltd.; 1990.
48. DeGroot MH. Optimal Statistical Decisions. McGraw-Hill; 1970.
49. Ferson S, Oberkampf WL, Ginzburg L. Validation of imprecise probability models. Sandia National Laboratories; 2008.
50. Beer M, Ferson S, Kreinovich V. Imprecise probabilities in engineering analysis. University of Texas at El Paso; 2013
51. Bernoulli D. Exposition of a new theory on the measurement of risk. In: Page A (editor). Utility Theory: A Book of Readings. John Wiley & Sons, Inc.; 1968. pp. 199–214.
52. Jolini S, Hazelrigg G.A. Differential utility: Accounting for correlation in performance among design alternatives. In: Proceedings of the ASME 2021 IDETC/CIE; August 17–19 2021.
53. Hazelrigg GA, Busch AC. Setting navigational performance standards for aircraft flying over the oceans. Transportation Research Part A: General. 1986; 20(5): 373–383.

Appendix

Flowchart for simulation of the M&M game

A flowchart for the simulation of the M&M game is shown in **Figure A1**.

The left column of this flowchart provides K simulations of the game. g are the guesses of other players and g^* is the winning guess of the other players. Thus, you win if your guess is between g^* and N. The values of g^* and N are saved for every simulation of the game. The right column computes the probability of winning the game as a function of your guess, n. As can be seen from this flowchart, the code necessary to compute the best guess, n, using a Monte Carlo simulation can be quite simple.

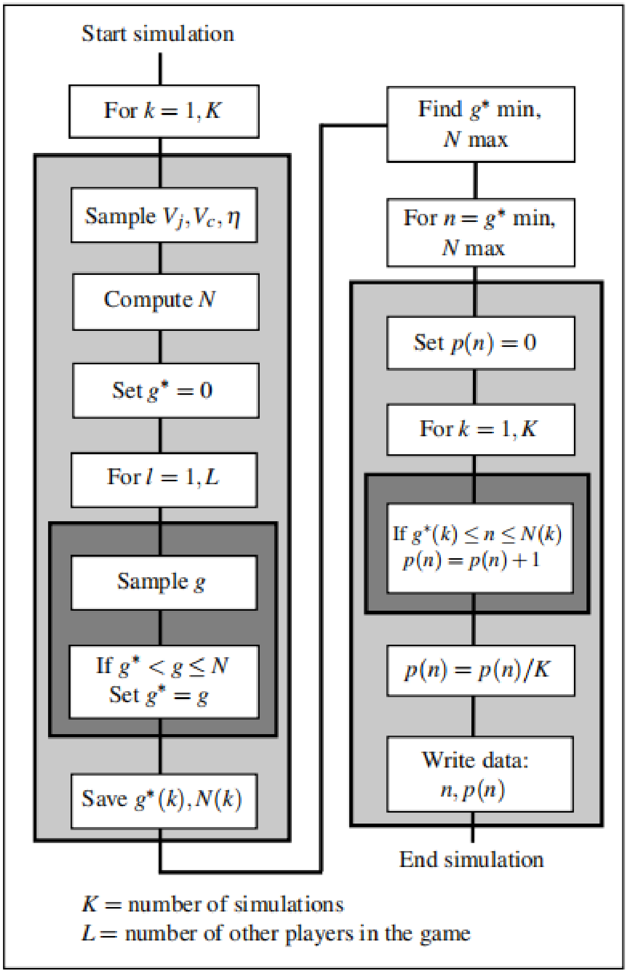

Figure A1. Flowchart for simulation of the M&M game.