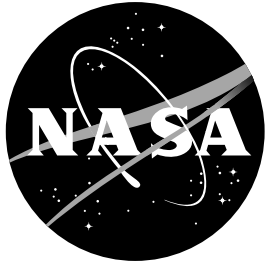


NASA/TM-2016-219341



Defining Baconian Probability for Use in Assurance Argumentation

*Patrick J. Graydon
Langley Research Center, Hampton, VA*

October 2016

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- **TECHNICAL MEMORANDUM.** Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- **CONTRACTOR REPORT.** Scientific and technical findings by NASA-sponsored contractors and grantees.

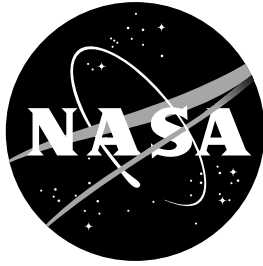
- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- **TECHNICAL TRANSLATION.** English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI Program, see the following:

- Access the NASA STI program home page at <http://www.sti.nasa.gov>
- E-mail your question to help@sti.nasa.gov
- Phone the NASA STI Information Desk at 757-864-9658
- Write to:
NASA STI Information Desk
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

NASA/TM-2016-219341



Defining Baconian Probability for Use in Assurance Argumentation

Patrick J. Graydon
Langley Research Center, Hampton, VA

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

October 2016

Acknowledgments

We thank the branch and directorate reviewers for their feedback on this work.

The use of trademarks or names of manufacturers in this report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Abstract

The use of assurance cases (e.g., safety cases) in certification raises questions about confidence in assurance argument claims. Some researchers propose to assess confidence in assurance cases using Baconian induction. That is, a writer or analyst (1) identifies defeaters that might rebut or undermine each proposition in the assurance argument and (2) determines whether each defeater can be dismissed or ignored and why. Some researchers also propose denoting confidence using the counts of defeaters identified and eliminated—which they call Baconian probability—and performing arithmetic on these measures. But Baconian probabilities were first defined as ordinal rankings which cannot be manipulated arithmetically. In this paper, we recount noteworthy definitions of Baconian induction, review proposals to assess confidence in assurance claims using Baconian probability, analyze how these comport with or diverge from the original definition, and make recommendations for future practice.

1 Introduction

The *safety case approach* has been used in some industries and regulatory domains for many years [1]. An organization using the approach adopts an appropriate safety management system, performs a hazard assessment, selects appropriate controls, and implements these. A written *safety case* documents the safety management system, hazards, controls, and evidence of the controls' adequacy [2]. The safety case combines safety evidence such as fault tree analysis results and test reports with an *safety argument*, typically defined as an “argument . . . that a system, service or organisation will operate as intended for a defined application in a defined environment” [3]. When similar cases are made for properties beyond safety, the general terms *assurance case* and *assurance argument* are used. A safety case might serve many purposes, one of which is to explain the safety rationale and evidence to an assessor who must decide whether the hazard controls and related evidence are adequate. Such use, which has analogues in security and other assurance matters, raises the question of assurance argument sufficiency. This question leads in turn to the concept of *confidence* (and its opposite, *uncertainty*) in the argument's claims.

In 1620, Francis Bacon proposed a scientific method that relies on what we now know as *Baconian induction*. Safety researchers propose using this induction to reason about the confidence justified by assurance cases [4–9]. They propose that analysts (a) identify doubts about safety claims (known variously as argument *defeaters* or *assurance deficits*) and (b) determine whether each can be dismissed or ignored and, if so, why. Some further propose denoting confidence using the counts of defeaters identified and eliminated, a metric they call *Baconian probability* and propose to manipulate arithmetically [4–6, 8, 9]. But Baconian probability was first formalized as a family of ordinal ranks, which cannot be manipulated arithmetically. This subtle difference has large implications for the validity of the proposed measure.

In Section 2, we summarize Bacon’s scientific method and noteworthy definitions of Baconian induction and Baconian probability that appear in the relevant foundational literature. In Section 3, we review current proposals to use Baconian induction to assess confidence in assurance claims. We show that two proposed uses of Baconian probability change its definition subtly but fundamentally from an ordinal rank to a cardinal measure. In Section 4, we analyze how the differences were key to the original concept’s validity and explore what remains to be shown about the validity of the new measure. In Section 5, we conclude by recommending (i) that the new measure be renamed to avoid confusion with the original and (ii) that the need for careful study of its validity be communicated clearly to researchers, regulators, and developers.

2 Background

In 1620, Bacon defined a scientific method that relies on a specific form of induction¹ [10, 11]. He was not the first to propose using induction: philosophers have defined and criticized forms of induction for millennia [12]. But Bacon’s induction is defined in terms of testing the impact of *relevant variables* rather than counting instances. Bacon did not define his form of induction precisely [10, 11]. But later philosophers defined a syntax for Baconian induction and proposed using Baconian probability to grade conclusions reached by chains of inference from evidence [13–17].

In this section, we summarize the background needed to understand Baconian induction and how it might be used to grade confidence in an assurance argument’s claims. We recount (1) Bacon’s scientific method, which relies upon a method of induction by consideration of relevant variables that Bacon does not define in detail; (2) Cohen’s proposed formalization of such a system of induction; and (3) Schum’s extension of Cohen’s system to chains of reasoning in applications beyond science.

2.1 Bacon’s Scientific Method

Baconian probability takes its name from Francis Bacon [10, 11]. His contribution to the subject was to define a procedure for deriving general rules of nature from scientific observations by using “true induction” [10].

A scientist using Bacon’s method to investigate a phenomenon compiles a set of tables and uses them to generate and assess hypotheses. The scientist begins by compiling a *table of instances*. Each instance is an observation of the occurrence of the phenomenon of interest, Bacon calls a *nature*. For example, Bacon’s example instances of *heat* include “the rays of the sun, especially in summer and at noon” and “lightning” [10]. The scientist then compiles a *table of divergency* comprising “instances where the nature of heat is absent but which are in other ways close to ones where it is present.” The scientist

¹ The term induction has different meanings in argumentation and mathematics. Here, we use it to refer to a form of argument in which conclusions are likely but not entailed.

finishes data collection by compiling a *table of degrees*. This table catalogues “instances in which the nature being investigated is found in different degrees, . . . either by comparing the amounts of it that a single thing has at different times or by comparing the amounts of it that different things have.”

The scientist is to use the tables to define hypotheses about natures causing or giving rise to other natures “and then, after a sufficient number of negatives, to reach a conclusion on the affirmative instances”:

The job of these three tables is . . . to present instances to the intellect. After the presentation has been made, induction itself must get to work. After looking at each and every instance we have to find a nature which

- is always present when the given nature . . . is present,
- is always absent when the given nature is absent,
- always increases or decreases with the given nature, and
- is a special case of a more general nature [10].

Bacon derides “the induction that proceeds by simply listing positive instances” for providing “conclusions on the basis of too few facts.” He proposes instead that the strength of support for a generalization grows with the number of relevant circumstances it has been shown to hold in, noting that a single counterexample disconfirms a hypothesis. Bacon defines heuristics for good source data in the form of *privileged instances* to be put into the tables. But while he proposes “to establish degrees of certainty,” the first volume of his *Novum Organum* does not define a mechanism for assessing or quantifying confidence. While Bacon planned to define induction more precisely in a later volume, he never did [12].

2.2 Cohen’s Eliminative Induction

In a series of works, Cohen developed both a scientific method and an accompanying syntax of induction that is variously called Baconian induction, *eliminative induction*, or *variative induction* [13–15]. He illustrates these by describing how von Frisch investigated a generalization about bees, namely that they discriminate between different colors. Von Frisch tested this hypothesis by testing whether bees would return to a transparent source of sugar placed on a piece of blue cardboard.

2.2.1 The Method of Relevant Variables

In Cohen’s *Method of Relevant Variables*, each field of enquiry has its own established list of relevant variables [14, §§42–45]. Each variable represents a type of circumstance that “suffices to falsify at least one generalization in that field.” For example, von Frisch investigated the hypothesis that bees discriminate between different colors by testing their ability to use color to return to a provided source of food. He manipulated the shape, odor, etc. of the

food source to investigate whether these other variables might better explain the bees' ability to return to the food source.

The list of variables is sorted in decreasing order of falsificatory potential based on “empirically influenced judgements of relative importance” [14, §45]. A corresponding list of tests t_1, t_2, \dots, t_n determines whether altering each variable or circumstance falsifies a proposed generalization. Each test also hierarchically includes the preceding tests. That is, it includes manipulations of both a new variable of interest and all of the variables manipulated by preceding tests. Thus, if a generalization passes test t_i and $j < i$, it must also pass test t_j .

The support that evidence gives to each generalization in a field can thus be ranked or graded according to the last test each passes. For example, suppose hypothesis H_1 passes test t_i but hypothesis H_2 does not. Because the tests are ordered and inclusive, it is not possible that a tested variable falsifies hypothesis H_1 but fails to falsify hypothesis H_2 . But there is at least one variable that falsifies hypothesis H_2 but not hypothesis H_1 . It can thus be said that the evidence supports hypothesis H_1 better than hypothesis H_2 .

Cohen denotes the evidential support for a generalization $i|n$, where t_i is the last test passed and n the number of tests. This figure is sometimes called the Baconian probability of the generalization on the given evidence.

The form of induction used in Cohen's and Bacon's scientific method is not the more familiar *enumerative induction*. Cohen calls his form of induction eliminative (or variative [15]) to distinguish it from this better-known alternative. While enumerative induction is concerned with how many tests a generalization passes, eliminative induction is concerned instead with how many types of circumstances it has been shown to hold in. Replication of a test might show that prior results were erroneous. If a generalization seemed to pass t_i but replication showed that it did not, the rank of evidential support for that generalization would fall to $i - 1$. But no amount of replication can show that a generalization holds in a new kind of circumstance. Thus, no amount of repetition could raise the rank of support to $i + 1$.

2.2.2 Grading Inductive Support for Statistical Hypotheses

The difference between Baconian induction and statistical testing (e.g., life testing to establish a component's failure rate) illustrates the nature of Baconian induction. This difference can be seen in Cohen's suggestion for grading hypotheses about statistics. Consider hypotheses of the form

$$p[S, R] = p \pm \varepsilon \tag{1}$$

i.e., the probability that any given R is an S is $p \pm \varepsilon$. Cohen asserts that each such hypothesis “has at least one equivalent that is a universally quantified conditional” [13, §13]. This might take the form “for any set V , if V is included in R , there is a set W such that V is included in W and W is co-extensive with R and $p[S, W] = p \pm \varepsilon$.” When testing this, “what is required is that there should be some set of objects or events that satisfies three conditions: it must be a sample of R on which an estimate $p[S, R] = p \pm \varepsilon$ may be based, it must

be selected under some particular manipulation of relevant variables, and its composition must be unbiased by other factors than variants of the variables being manipulated.” The inductive grading of the quantified conditional form

must not be conceived of as a measure of the evidential support for the hypothesis as a whole, but rather as an indication of the nature of the trials from which this-or-that grade of support is sought or attained. To have a higher coefficient of confidence for the same estimate we need a larger sample, and this may cost time, trouble and money. . . . A higher grade of support, on the other hand, may not be so readily available. Whether we can achieve it or not depends on the operation of the relevant natural variables as well as on the price we ourselves are prepared to pay for a performance of the appropriate test. The grade of inductive support attained by a hypothesis is determined solely by the extent to which it remains unfalsified under the manipulation of a cumulative hierarchy of relevant variables.

In this usage, the grade of inductive support is not a measure of the proportion of a sampled population that exhibits a characteristic. Nor does it measure the degree to which a measurement of that proportion reflects the effect of random chance on which members of a population were sampled. Instead, it reflects our knowledge about whether the sample was a sample of the population specified in the hypothesis or of a materially different subset of that population. The more the statistical hypothesis is tested and holds across subpopulations defined by relevant variables, the more confidence one should have in it.

2.2.3 An Ordinal Rank, Not a Cardinal Measure

Cohen’s formulation of induction provides a way to rank and compare the degree to which hypotheses within a specific class are supported by evidence. But Cohen insists that Baconian probabilities are ordinal rankings of evidential support, not cardinal measures of that support. That is, the inductive ranks of evidential support for each of a set of comparable hypothesis show which is best supported, which second best, and so on, but not by how much. Because the relevant variables—and the support function defined by the list of tests covering them—are specific to each class of hypotheses,

the support-functions for different categories of generalizations are largely incommensurable with one another. Zero and maximal values of these functions do have the same significance for each category, betokening no resistance and full resistance, respectively, to the falsifying effects of relevant variables. But intermediate values cannot be equated because of the differences in number, complexity, importance, etc. of the relevant variables in different series [14, §45].

Cohen considered potentially universal measures of inductive support such as “the numbers of appropriate trials—trials in variants of variables relevant to [a hypothesis]—that [the evidence] reports [the hypothesis] to have passed or failed” [13, §5]. He rejected that proposal for several reasons, two of which are grounds for despairing that any viable alternative could ever be found:

First, there is no reason to suppose that test-results on materially dissimilar hypotheses are commensurable. . . . The concept of a test articulated here is one that relates a test to a class of empirically discoverable variables that are relevant for any one of a class of materially similar hypotheses. This concept seems to allow no empirical basis for cross-field comparisons between materially dissimilar hypotheses. On what grounds would it legitimately be assumed that the interconnections of factors and characteristics prevailing in one field were precisely paralleled in another . . . which were not also grounds for taking hypotheses in these two fields to be materially similar? . . .

Thirdly, and more seriously, one test may be capable of giving greater support than another even when the number of variables manipulated is the same or smaller, because one of these variables is especially relevant—especially successful at falsifying hypotheses of the type in question. For example, mating/not-mating is perhaps a more relevant variable than temperature in relation to hypotheses about bird-plumage; and pregnant/not-pregnant is perhaps a less relevant variable than medical history in relation to hypotheses about drug-toxicity [13, §5].

2.3 Schum’s Extension to Chains of Reasoning

Schum proposes using Bayesian probability to rank evidential support for general hypotheses [16,17]. While Cohen focused on grading support for scientific hypotheses, Schum’s interest includes both *nonreplicable* events and *chains of reasoning*. Schum repeatedly uses the example of witness testimony in a legal proceeding.

2.3.1 The Witness Testimony Example

In Schum’s example, a witness W gives testimony E^* that event E occurred [17, §3.2.3]. One might challenge this testimony on at least three grounds:

- *Veracity*, i.e. the witness’s testimony E^* might differ from the witness’s belief about whether event E occurred
- *Objectivity*, i.e. the witness’s testimony E^* might be inconsistent with the witness’s observations related to event E
- *Observational sensitivity*, i.e. the witness’s observations might not accurately reflect whether event E occurred

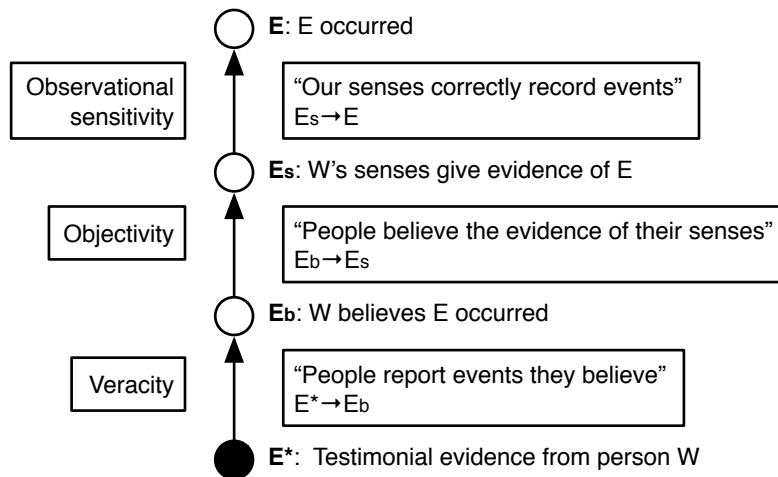


Figure 1: A synthesis of Figures 3.6 and 5.10 from [17]. The filled and open circles represent evidence and propositions, respectively. Connecting lines show inferences.

Finding these three issues to be distinct, Schum presents a three step argument from witness W 's testimony E^* to the proposition that event E occurred. Figure 1 depicts this argument. The first reasoning step uses a generalization about witness veracity to infer from witness W 's testimony E^* that the witness believes that event E occurred (E_b). The second reasoning step uses a generalization about witness objectivity to infer from witness W 's belief that the witness's senses detected that event E occurred (E_s). The final reasoning step uses a generalization about witnesses' observational sensitivity to infer from witness W 's having sensed event E that event E actually happened. Clearly these three generalizations do not apply with full force to all witness testimony.

2.3.2 Grading the Credibility of a Witness

While Cohen's work focused on single inferences, Schum proposes grading the force of multiple inferences through *detachment of belief* in the intermediate premises. Earlier argument steps are to be evaluated first and their conclusions detached—that is, taken as true at a given Baconian probability—when evaluating later argument steps. In the case of Schum's witness testimony example, an observer evaluating the strength of the testimonial evidence should begin by grading the force of the first inference:

Suppose that there is some number n_v of possible tests of this veracity generalization to see if it does in fact apply to W and his present testimony E^* Answers to questions such as these supply ancillary evidence about W 's veracity. Suppose we have asked i of these questions and that the answers are all favorable to the generalization [$E^* \rightarrow E_b$] so far as W is concerned. In Baconian terms, if we have confidence in these answers, we are entitled to *detach* a belief in E_b *at least* at grade level i/n_v [17, §5.5.3].

By grading the generalizations $E_b \rightarrow E_s$ and $E_s \rightarrow E$ in the same way, an observer can detach beliefs first in E_s and then in E . In Schum’s example, witness W passes j of n_o tests of objectivity and k of n_s tests of observational sensitivity, so the relevant Baconian probabilities are j/n_o and k/n_s .

2.3.3 Arithmetic on Baconian Probabilities?

Schum then notes that the tests not passed for each stage of the argument shown in Figure 1 represent a *ceteris paribus* (i.e., “all other things being equal”) assumption. He characterizes the strength of this assumption in terms of the numbers of questions identified and settled favorably:

A long-standing expectation is that inferences should be weakened as we add links in chains of reasoning. There is a definite Baconian counterpart to this expectation. The number of unanswered questions accumulates as we reason from one stage to another. The detachment of belief at any stage of an inference actually involves consideration of the total number of recognized questions that have not been answered at this stage and at stages lower in the chain of reasoning. In the example we have left unanswered $(n_v - i)$ at the first stage, $[(n_v - i) + (n_o - j)]$ at the second, and $[(n_v - i) + (n_o - j) + (n_s - k)]$ at the third. Thus, for example, our final detachment of a belief that E did occur, based on W ’s testimony, has a *ceteris paribus* assumption with content equal to the $[(n_v - i) + (n_o - j) + (n_s - k)]$ questions that have been left unanswered. This content indicates how complete has been our coverage of matters relevant to W ’s credibility. The more questions we leave unanswered about W ’s credibility, the less confident we can be in believing what he tells us [17, §5.5.3].

A slightly different version of this example appears in earlier work [16, §3.4.2]. In it, Schum defines the “inductive probability” that event E happened on testimony E^* and the detached beliefs in E_b at level i and E_s at level j as at least²

$$\frac{(i + j + k + 3)}{(n_v + n_o + n_s + 3)} \tag{2}$$

The constants in Equation 2 represent a benefit of the doubt given at each stage of the argument. Schum then observes that

in all of these formal expressions that the *difference* between numerator and denominator in a Baconian probabilistic expression always indicates the number of recognized relevant variables that have *not* been included in the test of credibility-related generalizations. Stated another way, the difference between numerator and

² In his earlier work, Schum uses the variables j , k , and m in place of i , j , and k in his later work and n_b in place of n_o . I present this example in terms of the variables used in the later version for the reader’s convenience.

denominator in a Baconian probability shows (i) the number of relevant credibility-related queries left unanswered, (ii) the content of one or more *ceteris paribus* assumptions, or (iii) the “weight” of the ancillary evidence resulting from the number of tests actually performed when this number is compared with the total number of tests that might have been performed.

A reader might take Schum to be defining Baconian probabilities in terms of numbers of unanswered questions and the Baconian probability of a multi-stage inference as the sum of unanswered questions at each stage. But such an interpretation would conflict with what Schum writes about the nature of Baconian probabilities elsewhere in the same works.

2.3.4 An Ordinal Rank, Not a Cardinal Measure

In several places, Schum reminds readers that Baconian probabilities are ordinal in nature and thus “cannot be meaningfully combined in any algebraic way” [16, §1.1, §3.4.1, §5.0], [17, §2.3.1, §§5.5–5.5.3, §5.6.3]. Relaxing Cohen’s assumption of inclusive tests, Schum writes that

the Baconian probability $B(H, E^*) \geq i/n$ means that the generalization licensing an inference of H from evidence E^* has been supported through level i in a sequence of n evidential tests involving variables believed to be relevant to the testing of this generalization [17, §5.5.2].

But, as in Cohen’s definition, the tests remain ordered. Baconian probability thus defines an ordinal rank of evidential support:

Grading the support [that] evidence provides hypotheses in eliminative and variative inference can only be performed in comparative terms (i.e., they cannot be combined algebraically), as these two properties of monadic and dyadic Baconian probabilities show:

- i. For monadic probabilities: $B(H_1) \geq B(H_2)$ or $B(H_2) \geq B(H_1)$.
For dyadic probabilities: $B(H_1, E_1^*) \geq B(H_2, E_2^*)$ or $B(H_2, E_2^*) \geq B(H_1, E_1^*)$.
- ii. For monadic probabilities: if $B(H_1) \geq B(H_2)$ and $B(H_2) \geq B(H_3)$, then $B(H_1) \geq B(H_3)$.
For dyadic probabilities: if $B(H_1, E_1^*) \geq B(H_2, E_2^*)$ and $B(H_2, E_2^*) \geq B(H_3, E_3^*)$, then $B(H_1, E_1^*) \geq B(H_3, E_3^*)$.

These two properties concern ordinal gradings (property i) that are also transitive in nature (property ii) [17, §5.5.2].

A reader might suppose that Schum both reports Cohen’s definition of Baconian probabilities as ordinal and proposes his own non-ordinal interpretation. But Schum writes that Cohen objected to interpreting Baconian probability as a number between 0 and 1 for three reasons:

First, it is clear that not all experimental tests are on the same par. In [Cohen's scientific method], the tests become successively more complex as well as more thorough. What this means, for example, is that the difference between $4/n$ and $3/n$ is not necessarily the same as the difference between $2/n$ and $1/n$. By itself this rules out [support for hypothesis H on evidence E] having interval or cardinal scale properties (and also ratio properties). Second, there is no obvious way to grade the degree of relevance of the tests being employed any more than there is to grade the relevance of any evidence. In the example the best you can do is to stand your tests in rank order in terms of their perceived importance. Third, there would be considerable difficulty in the meaning of i/n as a *fraction* when it is used grade the support provided to different hypotheses tested under possibly different circumstances. Suppose, for example, that you have a competitor who has produced her version S' of this system; she also performs n tests of S' . Suppose that your version S passes the first three tests out of n you performed and that her version S' passes 4 tests out of the n she performed. These results would not be comparable in any cardinal way if she ordered her eliminative tests differently or employed different tests [17, §5.5.1].

In neither work does Schum claim that these objections do not apply to his proposed use of Baconian probability. While he points out that achieving discipline-wide agreement on the ordering of tests for a class of generalizations might be difficult outside of science, he does not cite this as grounds for reimagining Baconian probability as a cardinal measure:

Experience in legal and other affairs supplies us with a list of relevant questions that might be asked regarding how well the above generalization holds in the present case of W and his testimony E^* The more of these tests that W passes, the more we are entitled to infer that this veracity generalization holds in the present instance of W and his testimony E^* But we cannot expect to manipulate relevant veracity-related variables in the same way that we manipulated the variables in the systematic testing of system S . In the first place, W has made a specific report about a unique or singular event that either happened or didn't happen; this is one reason why there are no veracity-related statistics. Second, there would almost certainly be argument about the relative importance of the six veracity questions as far as W is concerned. There might also be disagreement about whether W has actually passed or failed any of these veracity-related tests. Though the idea of performing an increasingly thorough factorial test disappears, the essential Baconian idea remains in the testing of generalizations such as the one we have examined. A generalization is supported to the extent that this generalization survives our best attempts to

show that it is invalid in the particular instance of concern. Such support can only be graded in ordinal terms [17, §5.5.1].

2.3.5 A Plausible Interpretation of Schum

Schum’s discussion of arithmetic on counts of questions asked and answered (as reported in Section 2.3.3) might be interpreted as a proposal to redefine Baconian probability as a cardinal measure of the number of identified questions that have been satisfactorily answered. But that interpretation conflicts with his insistence that Baconian probabilities are ordinal (as reported in Section 2.3.4). It also leaves readers to wonder why Schum does not address charges—Cohen’s and his own—that variables and questions do not have uniform importance.

Schum’s words can be interpreted in a way that avoids this problem. In this interpretation, (i) Schum proposes to use ordinal Baconian probabilities as Cohen defined them to rank and compare evidential support for general claims and (ii) his references to counts of questions identified and answered are meant to illustrate and justify the validity of that mechanism.

Suppose that jurists define (or a particular jurist defines) an ordered set of questions to be asked about the applicability of each of the three generalizations described in Section 2.3.2. These questions define an ordinal scale by which to rank similar claims made on similar evidence. By detaching claims used in further inference at a particular rank, Schum accumulates a vector of ranks that collectively grade evidential support for the conclusion of the final inference. Suppose that two claims, H_1 and H_2 , are supported by similar chains of reasoning, so that the same questions and scale apply to both. If the rank of claim H_1 ’s evidential support at each position in the vector is no less than the corresponding rank for H_2 , one could say that the evidence for H_1 supports H_1 at least as well as the evidence for H_2 supports H_2 .

Returning to Schum’s courtroom testimony example, suppose that witness W_1 testifies (E_1^*) that event E_1 happened and witness W_2 testifies (E_2^*) that event E_2 happened. In keeping with the notation used in Section 2.3.2, let i_1 stand for the number of the last veracity question asked of both witnesses that was satisfactorily answered in the case of witness W_1 , etc. If $B(E_i, E_i^*)$ stands for the grading of the support that witness W_i ’s testimony provides for the claim that event E_i occurred,

$$(i_1 \geq i_2 \wedge j_1 \geq j_2 \wedge k_1 \geq k_2) \rightarrow B(E_1, E_1^*) \geq B(E_2, E_2^*) \quad (3)$$

The number of veracity, objectivity, and observational sensitivity questions (n_v , n_o , and n_s) need not appear in Equation 3 because the same ranking scales, defined by the same questions sorted in the same orders, are used to evaluate the support provided by both witnesses’ testimony.

On this interpretation, Schum’s proposed use of Baconian probability does not provide a total ordering of evidential support for a claim through multiple reasoning steps, much less a means of comparing support across different types of claims. If different jurists assign different orders to the veracity-related questions, they might reach different conclusions about the relative strength of

two witnesses' testimonies. The proposal might be useful nevertheless, and it has the virtue of comporting with what both Cohen and Schum have written about the syntax and nature of Baconian probabilities.

Under this interpretation, Schum's discussion of the counts of questions identified and asked does not redefine Baconian probability. The discussion instead illustrates Baconian probability and defends Schum's proposed application of it. If each greater Baconian grade of support for a proposition did not represent greater knowledge about the proposition's truth, a skeptic might question the proposal's validity. That an increase in rank represents a decrease in the number of unanswered questions is a defense against this line of attack. On this interpretation, the apparent contradiction disappears: Baconian probabilities are ordinal ranks, not cardinal counts of unaddressed defeaters, but as the former rises, the latter falls as expected.

3 Proposals for the Use of Baconian Induction in Assurance Argumentation

Researchers have proposed using Baconian induction in the writing and assessment of assurance arguments. In this section, we review the most relevant proposals and assess how each comports with the definitions given in Section 2. While the first comports with the original concept of Baconian induction, the latter two subtly redefine Baconian probability as a cardinal measure rather than an ordinal rank.

3.1 Assured Safety Arguments

Hawkins et al. propose using eliminative induction as part of *assured safety arguments* [7]. Each assured safety argument comprises both a *main safety argument* and a *confidence argument*. The former records safety claims and their relationship to evidence. The latter "documents the reasons for having confidence in the safety argument" [7]. *Assurance claim points* in the safety argument identify the parts of that argument to which each part of the confidence argument applies. Figure 2 illustrates this use. Writers using the authors' argument patterns then justify confidence in an argument step by arguing the truth of three premises:

1. "Credible support exists" for the evidence or inference
2. The *assurance deficits* have been identified
3. The remaining deficits are acceptable

Each assurance deficit represents a "knowledge gap that prohibits perfect (total) confidence" [7]. The argument supporting the first premise gives reasons to accept the safety argument step in question. The argument supporting the second premise speaks to the adequacy of the process used to identify possible shortcomings in that rationale. And the argument supporting the third premise

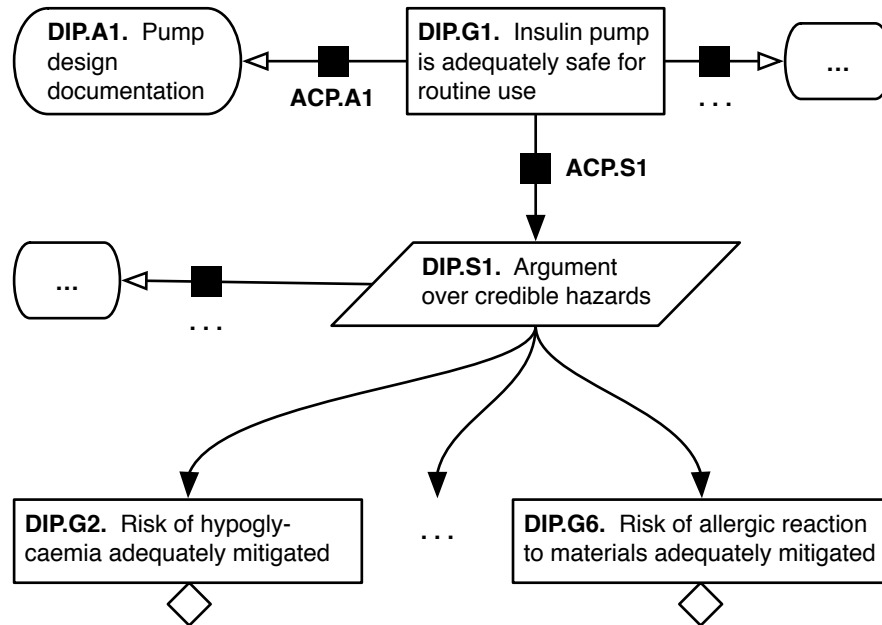


Figure 2: An example safety argument adapted from [7] and depicted in the Goal Structuring Notation [3] (with extensions). Assurance claim points such as *ACPS1* identify argument steps that are discussed in the confidence argument.

gives either reasons for believing either that (a) “significant counter evidence associated with [the] assurance deficit is sufficiently likely to exist” or (b) “the sensitivity of the safety argument to the assurance deficit is acceptably low.” Figure 2 illustrates the use of assurance claim points to identify steps in the main safety argument. Figure 3 presents a pattern for arguing confidence in inferences such as *ACPS1* in Figure 2.

Instantiations of the confidence argument shown in Figure 3 are an application of eliminative induction to assurance arguments. Bacon’s table of divergency and Cohen’s list of variables are means of identifying categories of circumstances in which a general rule of nature might not apply. In a confidence argument, “recognising assurance deficits . . . helps to identify the possible areas in the argument where counter-evidence *may* exist” [7]. Just as Cohen grades inductions according to entire categories of possible counterexamples tested, Hawkins et al. note that to identify assurance deficits is to “guid[e] . . . the otherwise boundless search for counterevidence.” The authors do not define a means of quantifying or ranking confidence in safety claims. However, they hypothesize that the proposed confidence argument patterns capture the details that are needed to support qualitative reasoning about confidence. A critical reader could search the confidence argument for evidence that its writers have overlooked an important assurance deficit or overestimated the degree to which evidence shows that an identified deficit is mitigated.

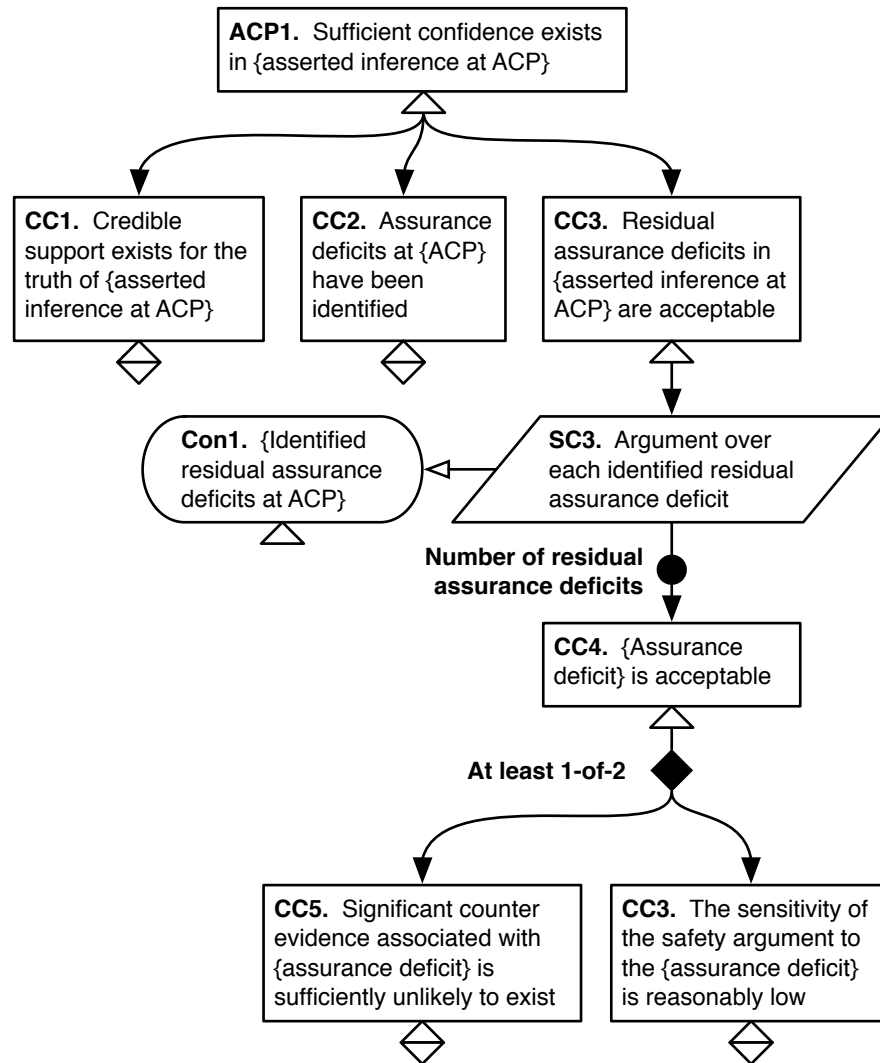


Figure 3: Confidence argument pattern adapted from [7]. A complete confidence argument would contain instantiations of this pattern for each assurance claim point like *ACP.S1* in Figure 2.

3.2 Eliminative Argumentation

Goodenough et al. propose an argumentation approach that includes both graphical *confidence maps* and Baconian probability [5, 6, 8, 9]. They name their complete approach *eliminative argumentation* to emphasize its focus on eliminative induction [9].

3.2.1 Confidence Maps

Like the argument pattern shown in Figure 3, a confidence map records possible reasons to doubt a safety claim and whether and how each is mitigated [6, 9].

Special symbols distinguish between three different kinds of argument defeater:

- *Rebutting defeaters*, which directly contradict claims
- *Undermining defeaters*, which represent doubts about evidence
- *Undercutting defeaters*, which signify doubts about inference rules

Unlike arguments in the Goal Structuring Notation, confidence maps do not end in evidence citations [3, 9]. Instead, confidence maps end when defeaters are marked either *OK* or *assumed OK*. When attached to an inference rule, the former “asserts that the rule has no undercutting defeaters because it is a tautology” [9]. Argument writers use the latter to signify having reached a point where “positing a new doubt seems unproductive.” A reviewer might challenge that assertion (or any other part of confidence map or assurance argument). Figure 4 gives an example the authors use to illustrate their proposal.

3.2.2 Baconian Probabilities in Eliminative Argumentation

Goodenough et al. build upon Schum’s summary of Cohen’s work on Baconian probabilities [9, 15, 17]. They explain that

For Cohen, the notion of “evidence” refers to the result of examining whether a hypothesis is favored when evaluated under various *conditions* that have the potential to cast doubt on the hypotheses (e.g., variations in temperature, humidity, shock, electromagnetic interference). He defines Baconian probability as $B(H, E) = i/n$, where E represents the number of tested conditions (n) as well as whether test results are deemed to favor or disfavor hypothesis H . In his formulation, results are available for all n test conditions, and i is the number of favorable results. $B(H, E) = i/n$ represents the tendency of the evidence to favor the hypothesis.

This description does not impose three conditions that are central to Cohen’s definition of Baconian induction, namely that:

1. Tests are defined specifically for classes of hypotheses of interest
2. Tests are ordered according to their relevance
3. Each test is inclusive of all lower-numbered tests

In eliminative argumentation, the Baconian probability of a claim in a confidence map is given as i/n , where n is the number of “defeaters at or nearest the leaves of the [argument] tree” (supporting that claim) and i is the number of these that have been eliminated (i.e., marked OK) [9]. For example, in Figure 4, the Baconian probability of claim C1.1 would be $i_{C1.1}/n_{C1.1}$, where $n_{C1.1}$ is the number of defeaters at leaves such as inference rule IR2.4 and $i_{C1.1}$ is the number of these that have been marked okay (also like inference rule IR2.4).

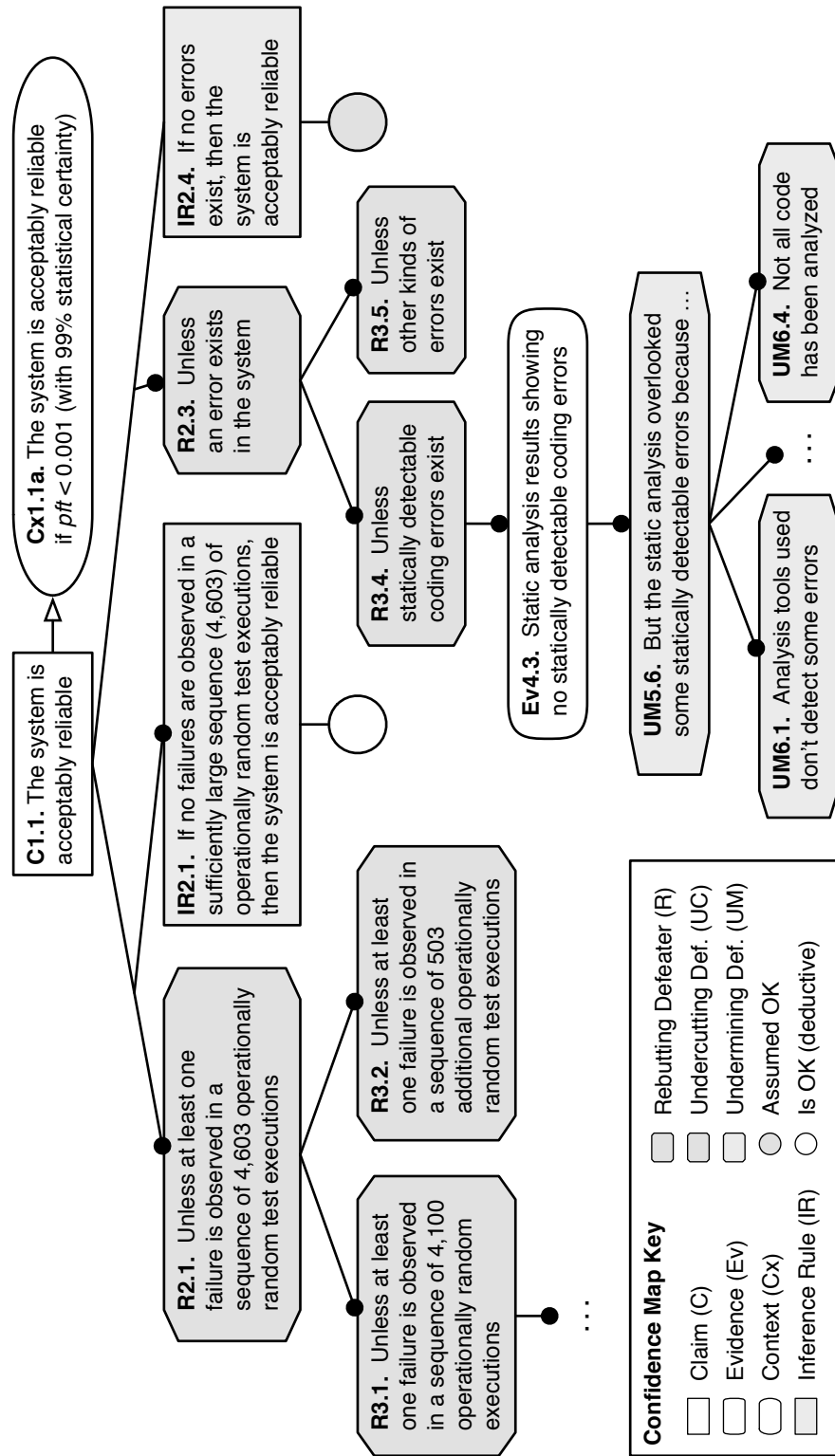


Figure 4: Example “multi-legged confidence map” adapted from [9, Fig. 24].

The practitioner is counseled that “it is best to focus not on the number of eliminated doubts or the fraction of eliminated doubts but instead to focus on the *residual* doubt—that is, the number of uneliminated doubts—because this represents the additional assurance work that is required to develop complete confidence in the top-level claim.” A prior report gives an example where two approaches to enumerating defeaters yield Baconian probabilities of $1/2$ and $3/4$ for the same claim and evidence [5]. That report cautions that the example

shows why Baconian probabilities cannot be compared directly and the care that must be taken when interpreting their meaning. Whether our confidence is expressed as $1/2$ or $3/4$, there is still one remaining doubt to be considered [5].

3.2.3 A Cardinal Measure, Not an Ordinal Rank

Like Cohen and Schum, the authors note that “in any set of defeaters, it is unlikely that they all seem equally important” [8]. If there are two defeaters and the first is more important than the second, then

if we are able to eliminate the first defeater and not the second, shouldn’t we have higher confidence in a claim of system safety than if we are able to eliminate the second defeater and not the first? Yet both situations would have Baconian probability $1/2$ [8].

Nevertheless, “incorporating a notion of relative importance of defeaters into our proposed grading of confidence . . . is not essential” because

Just as a system developer would not represent extremely unlikely and minimally impactful safety hazards in a safety case as a way of justifying an increase in confidence, under our framework a system developer would not take into consideration low impact defeaters and justify an increase in confidence by demonstrating their elimination [8].

In any case, “eliminating two out of three defeaters ($2/3$) provides more support for a claim than $1/3$ and less than $3/3$ ” [9]. By treating all defeaters as practically equivalent in consequence and performing arithmetic on them as described in Section 3.2.2, eliminative argumentation departs from Cohen and Schum by defining Baconian probability as a universal cardinal measure rather than as an ordinal rank of hypotheses in a given class.

The papers proposing eliminative induction do not define a fully realized framework for arithmetic on cardinal Baconian probabilities. Earlier work notes that such a probability “is neither a reducible fraction nor a fractional representation of a numerical value between zero and one” [5]. The most recent paper notes that “more work is needed to determine what methods of confidence calculation are most useful for given purposes and what practical effect various confidence numbers may have” [9].

3.3 Mapping to Beta Distributions

Duan et al. propose to assess confidence in assurance arguments using Baconian induction and quantify and visualize the Baconian probability of assurance claims using the beta distribution [4]. The beta distribution is characterized by two parameters, α and β . In the proposed quantification, “the number of doubts eliminated can be mapped directly onto the α parameter and the number of doubts remaining can be mapped directly onto the β parameter” [4]. Specifically, “assuming n is the Baconian numerator and d is the Baconian denominator,”

$$\alpha = n \tag{4}$$

and

$$\beta = d - n \tag{5}$$

Echoing Schum, this arrangement is said to make intuitive sense: “as the number of doubts eliminated grows, so does our confidence” [4].

Once reinterpreted as beta distributions, the Baconian probabilities of assurance claims can be used in further probability calculations:

We can then use properties of the beta distribution to calculate the uncertainty. Instead of simply summing up the Baconian probability values, ... we propose a weighting scheme with the beta distribution parameters. ... Uncertainty in the beta distribution can be calculated as the difference between the two inflection points of the curve [4].

Except for the special case where either α or β is equal to one, the authors define the confidence represented by a beta distribution as

$$\frac{2}{\alpha + \beta - 2} \sqrt{\frac{(\alpha - 1)(\beta - 1)}{\alpha + \beta - 3}} \tag{6}$$

The operators also propose to combine confidence in claims using a consensus operator that Jøsang defined for Shaferian belief frames [4, 18].

But what, to the authors, is a Baconian probability? The paper defines a Baconian probability as “a ratio of the number of doubts eliminated to the total number of doubts” [4]. It cautions that “this ratio is irreducible, as 4/5 would represent an entirely different confidence value than 8/10—the latter shows a higher confidence value and indicates that more doubts had been found and ultimately mitigated or eliminated.” But arithmetic is undefined on ordinal values. So, unlike Cohen or Schum, the paper must interpret Baconian probabilities as cardinal measures rather than ordinal ranks [14, 17]. This interpretation is corroborated by a citation to the earlier work of Goodenough et al. rather than to Cohen’s foundational texts [4, 5].

4 Baconian Induction in Assurance Arguments

As illustrated in Sections 2 and 3, some proposals to use Baconian probability to measure confidence in assurance cases define it differently from Cohen. Specifically, some safety researchers propose defining Baconian probability as a cardinal value related to the number of defeaters identified and addressed rather than as a family of ordinal ranks, each grading evidential support for a specific class of claims. This raises a question: do the difficulties that prompted Cohen to define Baconian probability in terms of incommensurable ordinal ranks—each grading evidential support for a specific kind of generalization—apply to assurance arguments? Might a cardinal, universal metric of unaddressed defeaters have utility even if it is subject to these difficulties?

4.1 Defeaters Vary in Importance

As discussed in Section 2.2.3, one of Cohen’s objections to treating Baconian probabilities as portable measures of evidential support was “the differences in number, complexity, importance, etc. of the relevant variables in different series” [14]. In assurance arguments, defeaters play the role of relevant variables. In some cases, well-known forms of inference might be associated with well-known lists of potential defeaters. In novel arguments or applications, new defeaters might apply. But what would be the impact if each defeater did not contribute an identical quantum to our confidence or uncertainty in the supported claim? Is there any evidence that defeaters have uniform importance?

If defeaters are not uniformly important, two arguments might provide different degrees of evidential support despite having the same number of unaddressed defeaters. This would occur whenever analysis of each revealed an equal number of unequally important defeaters. Similarly, two conclusions might have the same degree of evidential support yet have different cardinal scores. This would occur whenever an analyst identified a different number of defeaters covering the same overall sources of uncertainty.

An example suffices to both illustrate this possibility and demonstrate that assurance argument defeaters are not uniformly important. Suppose that the proposition in question is a claim about software behavior and the evidence comes from testing:

Claim C_1 : The software satisfies software safety requirement SSR1
(the definition of which is not germane).

Evidence E_1 : Software test results.

Suppose that two defeaters for the inference from E_1 to C_1 are identified:

Defeater D_1 : The software test results labeled as evidence E_1 were obtained by testing the wrong version of the software.

Defeater D_2 : The test report labeled as evidence E_1 might have been corrupted by random bit flips in the memory used by the word processing software with which it was prepared.

While the possibility of the bit flip at issue in defeater D_2 cannot be ruled out, the possibility of it leading to an undetected error in the test report is so remote that the defeater might strike some readers as absurd. But it is a potential defeater despite being obviously less important than defeater D_1 . To represent both defeaters as a single quantum on a cardinal scale of confidence or uncertainty is to conflate defeaters of very different importance, undermining the validity of comparisons of confidence or uncertainty based on that scale.

Goodenough et al. assume that developers will not deliberately choose the scope of defeaters so as to present a misleading indication of confidence [9, §3.2]. But this assumption is not sufficient to ensure that each defeater represents an identical quantum of confidence or uncertainty: two analysts might, with no intention to deceive, identify different-sized sets of defeaters that represent the same potential reasons for doubt. For example, Goodenough et al. give an example in which an analyst might choose to represent the same uncertainty arising from incomplete test coverage as either $1/2$ or $3/4$ depending on whether the analyst counted paths or basic blocks [5, §5.4].

One might hypothesize that, despite the existence of examples of defeaters that vary substantially in importance, *most* relevant assurance argument defeaters are similar enough in importance that any errors in assessment caused by their differences can be safely overlooked. If this hypothesis is true, the proposed cardinal measure might be useful despite its imperfections. But this hypothesis would require substantial empirical validation. No such empirical validation has yet been reported in the relevant literature.

4.2 The Order of Defeaters Matters

Cohen advised sorting variables in decreasing order of falsificatory potential based on “empirically influenced judgements of relative importance” [14, §45]. Perfect achievement of this order is not a prerequisite for defining a valid rank. Rather, approximating the recommended order is a matter of optimizing the cost of hypothesis testing: the closer that a test set defined for a given class of hypotheses achieves this order, the fewer tests will be needed on average to assess a hypothesis from that class.

As Schum points out, there might not be universal agreement about the order of importance of a set of questions about witness veracity [17, §5.5.1]. And as Cohen observed and Schum reiterated, there are situations such as courtroom testimony in which one does not have the luxury of testing variables in a specific order [14, §69] [16, §3.2] [17, §5.1.1].

In safety engineering, it is not clear that there is always an empirical basis for determining which of two defeaters is most important. Moreover, forcing analysts to consider and assess defeaters in a specific order would be a substantial imposition. Does this make it impossible to rank defeaters to a particular kind of inference used in assurance arguments? Is it still necessary to establish such a rank?

No and yes, respectively. While perfect order is unnecessary and while tests might not be conducted in a specific order, the existence of *some* nominal

order for tests of scientific hypothesis is key to the validity of the inductive ranking mechanism. If tests are ordered and hierarchically inclusive, rank $i + 1$ always indicates greater evidential support than rank i even if the difference between ranks i and $i + 1$ is not the same as that between ranks $i + 1$ and $i + 2$. The order of the tests makes comparisons of rank valid.

The same principle applies to Baconian induction in assurance arguments. It isn't necessary to perform the assessments of whether each defeater holds in a certain order. But if defeaters might not represent uniform quanta of confidence and do not have a nominal order, then greater rank of evidential support might not imply that greater confidence is warranted and vice-versa.

That some defeaters differ in importance is sufficient to show that relaxing the requirement for a nominal order of tests creates at least some cases where the order of assessed values misrepresents the strength of the assessed arguments. One might nevertheless hypothesize that defeaters are usually close enough in importance that such difficulties do not pose practical problems. But that hypothesis does not yet have any empirical support.

4.3 Ranks Are Incomparable Across Inference Types

Cohen's Baconian probabilities (except for $0/n$ and n/n) are incommensurable across types of hypotheses as described in Section 2.2.3. Schum described how one might, in some cases, compare the grade of support for identical-length chains of similar inferences as discussed in Section 2.3. But Schum's mechanism does not permit comparing the support provided by dissimilar argument structures. Is it possible to compute a Baconian probability that grades support for an arbitrary assurance argument such that the grade could be compared to either a threshold or the grade of an alternative argument?

No. In Cohen's formulation, different relevant variables might apply to different kinds of generalizations. Since different questions might have different significance, knowing that a generalization passed test t_i of n tests, where $n > i$, does not facilitate comparison of the degree of evidential support except to that for another hypothesis tested using the same tests.

Knowing that an assurance argument inference addresses the first i of an ordered set of n defeaters, where $n > i$, permits comparing the strength of that inference to another inference to which the same set of defeaters also applies. But, as discussed in Section 4.1, defeaters might not represent uniform quanta of confidence or uncertainty. Thus one might have addressed the first i of n defeaters for one inference, and the first i of a different ordered set of n defeaters for another, even if the actual uncertainty in the conclusions differs.

That some defeaters differ in importance is sufficient to show that there are at least some pairs of inference types such that Baconian probabilities representing the degree of support for each would be incommensurate (except for the limiting cases $0/n$ and n/n). Again, one might hypothesize that defeaters are usually close enough in importance that such difficulties are not problematic in practice. But that hypothesis does not yet have any empirical support.

4.4 Is Unaddressed Defeater Count Useful As a Metric?

As discussed in Section 3, some safety researchers define Baconian probability in a way that differs subtly but fundamentally from the original definition. That is, they use the term to refer to a universal cardinal metric related to the number of unaddressed defeaters rather than to a series of inference-type-specific ordinal ranks. As discussed in Sections 4.1–4.3, this difference relaxes conditions and restrictions that are key to the original version’s validity. But might the resulting metric nevertheless be usefully applied to assurance argumentation?

Possibly. But to establish such utility, researchers would need to (1) identify a (desirable) value that it brings, (2) demonstrate that use of the proposed metric brings that value, and (3) show that the value outweighs any harms that use of the metric also brings (e.g., the potential for error when a larger value actually stands for less confidence). We are not aware of any published studies that address either of the latter two needs. If such research is attempted, it would be helpful if the proposed metric were called by a new name to avoid confusion with the original definition of Baconian probability.

5 Conclusion

Baconian induction and probability differ from both induction by number of instances and mathematical induction. Bacon proposed to define and assess scientific hypotheses according to the number of relevant circumstances each holds in. Cohen formalized Bacon’s induction and defined a syntax and rules for Baconian probability. Schum proposed applying Cohen’s Baconian probability to the grading of evidential support for general claims supported by chains of inference.

Researchers have proposed applying these concepts to the representation and assessment of assurance arguments. All such proposals replace tests of relevant scientific variables with assessments of identified potential argument defeaters. One notable proposal stops at representing both (a) the list of defeaters and (b) the reason for thinking that each of these has been adequately addressed or can be safely ignored. Another includes a more syntactically rich notation for representing that information. This proposal, like a more recent third proposal, embodies a subtle but foundational redefinition of Baconian probability. While Cohen and Schum define Baconian probability in terms of ordinal ranks—each applicable to a specific class of hypotheses or inference types—these proposals define it as a universal, cardinal measure.

An ordinal rank specific to a type of inference enables comparing the degree of support offered by two alternative assurance argument steps using that type of inference. A universal, cardinal measure might offer more value, e.g. the ability to compare support for dissimilar types of inference or even a measure of the support provided by a complete argument. But Cohen’s definition of Baconian probability includes conditions that make rank comparison sound despite the non-uniform importance of relative variables. These conditions would likewise keep comparisons of the strength of similar assurance argument steps

from being unsound on account of differences in defeaters' importance. But proposed cardinal measures related to the number of defeaters identified and addressed relax these conditions. Given the non-uniformity of assurance argument defeaters, one argument might measure better than another when its support is in fact weaker.

Cohen's Baconian probabilities are ordinal ranks, not cardinal measures. Because this distinction is crucial to their validity, we recommend that proposed measures related to the number of defeaters identified and addressed be given a name other than Baconian probability to distinguish them from Cohen's ranks. While the non-uniformity of assurance argument defeaters threatens the validity of the proposed metrics, an imperfect metric might be useful if it is used with due regard for its potential to mislead. Such utility is an empirical question of both the value the metric might bring and the potential for harm due to misinterpretation. Neither that value nor that harm has been studied appropriately. Because the utility of the proposed measures remains in doubt, we recommend that future writing about these measures make clear to readers that their validity has not yet been empirically established.

References

1. U.S. Chemical Safety and Hazard Investigation Board: Chevron Richmond Refinery Pipe Rupture and Fire. Regulatory Report 2012-03-I-CA, CSB, Oct. 2014. URL <http://www.csb.gov/chevron-refinery-fire/>.
2. A-P-T Research, Inc.: Safety Case Workshop. Electronic document, 14–15 Jan. 2014. URL http://www.ap-t-research.com/news/2014-01-15_SafetyCaseWorkshop/T-13-00600%20Safety%20Case%20Workshop%20Findings.pdf.
3. Attwood, K.; et al.: *GSN Community Standard Version 1*. Origin Consulting Limited, York, UK, Nov. 2011. URL http://www.goalstructuringnotation.info/documents/GSN_Standard.pdf.
4. Duan, L.; Rayadurgam, S.; Heimdahl, M.; Sokolsky, O.; and Lee, I.: Representation of Confidence in Assurance Cases using the Beta Distribution. *Proceedings of the IEEE 17th International Symposium on High Assurance Systems Engineering (HASE)*, 2016. URL <http://dx.doi.org/10.1109/HASE.2016.52>.
5. Goodenough, J. B.; Winstock, C. B.; and Klein, A. Z.: Toward a Theory of Assurance Case Confidence. Technical Report CMU/SEI-2012-TR-002, Software Engineering Institute, Pittsburgh, PA, USA, Sept. 2012. URL <http://www.sei.cmu.edu/reports/12tr002.pdf>.

6. Goodenough, J. B.; Weinstock, C. B.; and Klein, A. Z.: Elimina-
tive Induction: A Basis for Arguing System Confidence. *Pro-
ceedings of the International Conference on Software Engineer-
ing (ICSE)*, San Francisco, CA, USA, May 2013, pp. 1161–
1164. URL [http://www.sei.cmu.edu/library/assets/
whitepapers/icse13nier-id10-p-15833-preprint.pdf](http://www.sei.cmu.edu/library/assets/whitepapers/icse13nier-id10-p-15833-preprint.pdf).
7. Hawkins, R.; Kelly, T.; Knight, J.; and Graydon, P.: A New Ap-
proach to Creating Clear Safety Arguments. *Advances in Systems
Safety: Proceedings of the 19th Safety-Critical Systems Symposium (SSS)*,
C. Dale and T. Anderson, eds., Southampton, UK, Feb. 2011, pp. 3–23.
URL [http://www.cs.virginia.edu/~jck/publications/
SSS.2011.safety.cases.pdf](http://www.cs.virginia.edu/~jck/publications/SSS.2011.safety.cases.pdf).
8. Weinstock, C. B.; Goodenough, J. B.; and Klein, A. Z.: Mea-
suring Assurance Case Confidence using Baconian Probabilities. *Pro-
ceedings of the 1st International Workshop on Assurance Cases for
Software-Intensive Systems (ASSURE)*, San Francisco, CA, USA, May
2013. URL [http://www.sei.cmu.edu/library/abstracts/
whitepapers/weinstock-icse2013-baconian.cfm](http://www.sei.cmu.edu/library/abstracts/whitepapers/weinstock-icse2013-baconian.cfm).
9. Goodenough, J. B.; Weinstock, C. B.; and Klein, A. Z.: Elimina-
tive Argumentation: A Basis for Arguing Confidence in Sys-
tem Properties. Technical Report CMU/SEI-2015-TR-005, Soft-
ware Engineering Institute, Pittsburgh, PA, USA, Feb. 2015.
URL [http://resources.sei.cmu.edu/asset_files/
TechnicalReport/2015_005_001_434813.pdf](http://resources.sei.cmu.edu/asset_files/TechnicalReport/2015_005_001_434813.pdf).
10. Bacon, F.: The New Organon or: True Directions Concerning the Interpre-
tation of Nature. Electronic document, May 2007. URL [http://www.
earlymoderntexts.com/assets/pdfs/bacon1620.pdf](http://www.earlymoderntexts.com/assets/pdfs/bacon1620.pdf),
translation of 1620 original by Jonathan Bennett.
11. Cohen, L. J.: Some Historical Remarks on the Baconian Conception of
Probability. *Journal of the History of Ideas*, vol. 41, no. 2, Apr.–June
1980, pp. 219–231. URL [http://dx.doi.org/10.1007/978-
94-017-2020-5_16](http://dx.doi.org/10.1007/978-94-017-2020-5_16).
12. Milton, J. R.: Induction before Hume. *Inductive Logic*, D. M.
Gabbay, S. Hartmann, J. Woods, and P. Thagard, eds., El-
sevier B.V., vol. 10 of *Handbook of the History of Logic*,
2011. URL [http://www.stephanhartmann.org/wp-
content/uploads/2016/02/HHL10_Milton.pdf](http://www.stephanhartmann.org/wp-content/uploads/2016/02/HHL10_Milton.pdf).
13. Cohen, L. J.: *The Implications of Induction*. University Paperbacks,
Methuen & Co., Ltd., London, UK, 1973.
14. Cohen, L. J.: *The Probable and the Provable*, vol. 10 of *Clarendon Library
of Logic and Philosophy*. Clarendon Press, Oxford, UK, 1977.

15. Cohen, L. J.: *An Introduction to the Philosophy of Induction and Probability*. Clarendon, Oxford, 1989.
16. Schum, D.: Jonathan Cohen and Thomas Bayes on the Analysis of Chains of Reasoning. *Probability and Rationality: Studies on L. Jonathan Cohen's Philosophy of Science*, E. Eells and T. Maruszewski, eds., Rodopi Press, Amsterdam, The Netherlands, vol. 21 of *Poznan Studies in the Philosophy of Sciences and the Humanities*, 1991, pp. 99–145.
17. Schum, D. A.: *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press, Feb. 2001.
18. Jøsang, A.: A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 3, June 2001, pp. 279–311. URL <http://eprints.qut.edu.au/7204/1/Jos2001-IJUFKS.pdf>.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) 01-10-2016		2. REPORT TYPE Technical Memorandum		3. DATES COVERED (From - To) June–October 2016	
4. TITLE AND SUBTITLE Defining Baconian Probability for Use in Assurance Argumentation				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Graydon, Patrick J.				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER 999182.02.50.07.02	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Langley Research Center Hampton, Virginia 23681-2199				8. PERFORMING ORGANIZATION REPORT NUMBER L-20749	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001				10. SPONSOR/MONITOR'S ACRONYM(S) NASA	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) NASA/TM-2016-219341	
12. DISTRIBUTION/AVAILABILITY STATEMENT Unclassified-Unlimited Subject Category 03 Availability: NASA CASI (443) 757-5802					
13. SUPPLEMENTARY NOTES An electronic version can be found at http://ntrs.nasa.gov .					
14. ABSTRACT The use of assurance cases (e.g., safety cases) in certification raises questions about confidence in assurance argument claims. Some researchers propose to assess confidence in assurance cases using Baconian induction. That is, a writer or analyst (1) identifies defeaters that might rebut or undermine each proposition in the assurance argument and (2) determines whether each defeater can be dismissed or ignored and why. Some researchers also propose denoting confidence using the counts of defeaters identified and eliminated—which they call Baconian probability—and performing arithmetic on these measures. But Baconian probabilities were first defined as ordinal rankings which cannot be manipulated arithmetically. In this paper, we recount noteworthy definitions of Baconian induction, review proposals to assess confidence in assurance claims using Baconian probability, analyze how these comport with or diverge from the original definition, and make recommendations for future practice.					
15. SUBJECT TERMS safety case, assurance argument, confidence, Baconian probability, eliminative induction					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk (email: help@sti.nasa.gov)
U	U	U	UU	30	19b. TELEPHONE NUMBER (Include area code) (443) 757-5802