Measuring Assurance Case Confidence using Baconian Probabilities

Charles B. Weinstock, John B. Goodenough, Ari Z. Klein
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA, USA
weinstock|jbg|azklein@sei.cmu.edu

***Abstract*— The basis for assessing the validity of an assurance case is an active area of study. In this paper, we discuss how to assess confidence in a case by considering the doubts eliminated by the claims and evidence in a case. This is an application of eliminative induction and the notion of Baconian probability as put forward by L. Jonathan Cohen.**

***Index Terms*—assurance case, eliminative induction, defeasible reasoning, safety case, Baconian probability, probability**

I. Introduction

Interest and use of assurance cases [1] [2] is increasing in both the United States and elsewhere [3]. The US Food and Drug Administration has taken steps towards requiring their use when a device manufacturer submits a medical device for approval [4]. In Europe, the use of the safety case (an assurance case for safety) is a well-established practice — required in some industries [5]. However, given an assurance case, how do you achieve confidence in its argument? This is a classic philosophical problem: determining the basis for belief (becoming confident) in a hypothesis when it is impossible to examine every possible circumstance covered by the hypothesis.

The issue of how to evaluate assurance case validity is gaining more attention. Hawkins et al. [6] discuss the use of *assurance claim points* as a means of identifying and mitigating *assurance deficits*. Lin and Kelly [7] discuss ways evidence can be faulty. Wassyng [8] notes the need for more work on confidence evaluation and suggests greater use of objective, prescriptive standards as an approach. Bloomfield et al. [9] [10] [11] have investigated a Bayesian approach.

Our solution [12] is to use an argumentation theory based upon defeasible reasoning, eliminative induction, and Baconian probability to help identify sources and amounts of confidence in a claim. In this brief paper we present the basics of the theory and explore more thoroughly the rationale for our use of eliminative induction and Baconian probability. In a companion paper [13], we present a notation based on our approach and provide a brief example.

II. Defeasible Reasoning and Eliminative Induction

In the argumentation literature, much of the existing work related to notions of confidence has been concerned with weighing which hypothesis or conclusion in a given set of possibilities is most strongly inferable from some pool of evidence [14]. For example, from a legal perspective, one might be interested in inferring which suspect, out of several potential criminals, robbed the bank, given evidence of where the suspects said they were at certain times, witness testimonials, and video surveillance. From an artificial intelligence perspective, one might be interested in how an artificial agent revises or maintains its belief in a reasonably derived conclusion as additional information becomes available.

The goal of these kinds of argumentation is to choose the hypothesis or conclusion, among several potential ones, that is the most acceptable or believable on the basis of the available evidence. Such argumentation theories are not directly relevant to assurance cases. In assurance case argumentation, rather than laying out several possible claims and searching for which is best supported by our existing collection of information, we start by asserting a single claim whose truth we wish to determine, and we gather evidence in support of that single claim. This difference means that certain complications that are important in argumentation theory (e.g., the interplay of argument and counter-argument) do not need to be addressed for assurance cases. Of course, other issues are directly relevant; for example, since arguments are based on a hierarchy of claims, we are interested in understanding how confidence in higher level assurance case claims is affected by confidence in lower level claims and information.

Although the assurance case argumentation goal is different from that with which much argumentation theory is concerned, we can nonetheless draw on existing argumentation theories for a theory of assurance case confidence. In particular, we ground our theory on the intersection of two frameworks of reasoning—defeasible reasoning and eliminative induction. Taken together, these frameworks provide both a conceptual vocabulary for identifying information that is relevant to measuring claim confidence and a basic method for justifying an increase in confidence.

Analyzing the soundness of a claim in terms of doubts identified and eliminated is an application of *eliminative induction,* a reasoning approach first articulated by Sir Francis Bacon [14]. In this approach, one identifies a variety of different reasons for doubting the validity of a hypothesis (or claim). As we eliminate such doubts, our confidence in the validity of the hypothesis necessarily increases.

Eliminative induction stands in contrast to *enumerative induction* in which confidence is built by finding increasing numbers of confirming instances, e.g., the more tests that run successfully, the more confident we might feel that a system is correct, or safe. But in the eliminative induction approach, unless tests eliminate some doubt about an argument, we don't know why their success should increase our confidence in a system property.

Eliminative induction depends on having a method for generating sets of doubts. When we consider assurance cases, notions from *defeasible reasoning* provide such a method. In defeasible reasoning, any conclusion is tentative because additional information can be used to attack an argument. In the defeasible reasoning literature [15] [16], the ways of attacking an argument are called defeaters. There are only three types of defeaters: rebutting, undermining, and undercutting. A *rebutting* defeater provides a counter-example to a claim. An *undermining* defeater raises doubts about the validity of evidence. An *undercutting* defeater specifies circumstances under which the conclusion is in doubt when the premises of an inference are true.

TABLE I. KINDS OF DEFEATERS

Defeater	Attacks	Form	Mitigation
Rebutting	Claim	R, so claim C is false	Look for counter-examples and why they can't occur
Undercutting	Inference rule	U, so claim C can be true or false	Look for conditions under which the rule is not valid and why those conditions don't hold
Undermining	Evidence	M, so premise E is false	Look for reasons the evidence might be invalid and show those conditions don't hold

As a simple example, we might argue "Tweety can fly because Tweety is a bird." This argument is supported by the inference rule, "If X is a bird, X can fly." But suppose Tweety is actually a penguin (a rebutting defeater), or that Tweety is actually a bat (an undermining defeater attacking the premise that Tweety is a bird), or that although Tweety is a member of a bird species that flies, Tweety is a juvenile (an undercutting defeater because the premise is true but the conclusion is uncertain). In each of these cases, we have raised doubts about the validity of the argument and the correctness of its conclusion. Identifying defeaters and removing them (by showing that they are false) is the essence of our approach to building and assessing confidence.

Defeasible reasoning provides a framework for identifying the reasons for doubt used in eliminative induction (namely, the defeaters), and eliminative induction provides a basis for justifying confidence in a claim, namely, by demonstrating that the defeaters are not valid.

III. Baconian Probability and Confidence

Our framework for confidence is rooted in eliminative induction and the Baconian system of probabilities as elaborated by Cohen [17]. Cohen's view is that probabilities are a way of "grading certainty" or "levels of support" for a hypothesis. His viewpoint incorporates Pascalian, Bayesian, and Baconian views of probability — all presented as different ways of grading certainty (or, as we call it, "confidence") in some hypothesis.

In eliminative induction, we identify different possibilities for doubting the truth of a hypothesis, and then we gather evidence or perform analyses that eliminate each of these possibilities. Each eliminated possibility removes a reason for doubt and thereby increases our confidence in the hypothesis. In our adaptation of Cohen's formulation, if there are n possibilities for doubt and i of these have been eliminated, $i|n$ is the Baconian probability, B(H), that hypothesis H is true.[1] The uneliminated doubts, $n - i$, constitute *residual doubt* in the truth of the hypothesis. (In the context of assurance cases, the hypothesis is a claim and the only possibilities for doubt are captured in the defeaters associated with the argument supporting the claim.)

To be more specific, consider an assurance case claim, C, and suppose we have identified n defeaters that cast doubt on its truth. As long as C withstands all n defeaters (i.e., as long as all defeaters are shown to be false), we have no reason for doubting that C is true; there is no residual doubt. In this case, $i = n$ and therefore, B(C) = $n|n$, so we have "total (Baconian) confidence" in C. On the other hand, in the case where B(C) = $0|n$, we have no confidence in C. In eliminative induction, "no confidence" does not mean that C has been disproven; it means we have not yet eliminated any doubts relevant to C.

One can also view Baconian probabilities as measures of the amount of information we have that is relevant to removing reasons for doubt. B(C) = $0|n$ means no information is available, and B(C) = $n|n$ means we have all the information needed to remove all doubts about the truth of C.

For a given claim, as evidence is developed, doubts are eliminated, and the Baconian probability of the claim increases. For example, for a given claim C, and set of doubts, $n,$ confidence in the truth of C increases as doubts are eliminated and B(C) increases towards $n|n$.

Baconian probabilities in general are not directly comparable: we cannot say that B(C) = $6|8$ implies less confidence in C than B(C) = $7|9$. In either case, two doubts remain. However, in some cases, it seems intuitively reasonable to say that for a given argument, if another doubt is identified and eliminated, the revised argument provides more support for C than the old. For example, if a new hazard is identified and then eliminated, we might say that Baconian confidence has increased from $n|n$ to $n+1|n+1$. But from a practical viewpoint in assessing assurance case confidence, residual doubt is more important than the total number of identified doubts because residual doubts represent the amount of work needed to develop total confidence in a claim.

IV. Concerns With This Approach to Confidence

Various concerns have been raised about the theoretical and practical utility of this approach. We discuss five of them here:

[1] We have modified Cohen's notation. Cohen would say i/n, but i/n is not to be interpreted as a fraction, so $i|n$ is less misleading.

- What if a relevant defeater has not been identified?
- What if a defeater cannot be completely eliminated?
- Surely not all defeaters are of equal importance. How is this handled?
- Baconian probability (eliminative induction) seems rather weak in contrast to Bayesian or Pascalian probability (enumerative induction). What is being gained (and lost) with this approach?
- The potential number of defeaters seems incredibly large for a real system. Is this approach practical?

A. Unidentified Defeaters

Since measuring confidence using eliminative induction is based on identifying and eliminating reasons for doubting the truth of a claim, failure to identify some relevant reasons for doubt would seem to give an inflated sense of confidence.

There are several answers to this concern. First, the inability to identify some defeaters is itself a reason for doubt that needs to be recognized in a case. For example, consider a claim that a system is safe. Counter-examples (rebutting defeaters) for such a claim would include possible hazards. Typically a case will argue that each identified hazard is eliminated or adequately mitigated. The implicit inference rule in such an argument is "If all hazards are eliminated/mitigated, the system is safe". An undercutting defeater for this inference is "Not all hazards have been identified". Therefore, to have full confidence, one must eliminate this undercutting defeater, e.g., with a claim that all hazards have, in fact, been identified. One must then produce an argument showing why one has confidence that all relevant hazards have been identified. Such an argument might, for example, rely on evidence that an appropriate and adequate hazard analysis has been done. If (as is usually the case in practice) one concludes that an appropriate analysis has been done to identify a particular set of hazards that need to be considered, then this undercutting defeater will have been eliminated. (Of course, all the hazards identified in the hazard analysis must actually be addressed by the case.)

So one answer to the concern about unidentified defeaters is that an adequate case will have considered this possibility and built an argument asserting that all relevant defeaters have been identified. Nonetheless, an inadequately prepared case will almost certainly have failed to identify some relevant defeaters. An assessment process for reviewing a case will need to take this possibility into account. What our approach offers is a consistent and theoretically exhaustive method for identifying sources of doubt, namely, by examining every inference rule for missing undercutting defeaters, every item of evidence for missing undermining defeaters, and every claim for possible counter-examples. We expect that this framework will help trained individuals to, in practice, reach agreement that all relevant defeaters have been identified; of course, this remains to be demonstrated.

Finally, eliminative induction and defeasible reasoning concepts are ways of thinking about and explaining why one should have confidence in a case, or a claim. The concepts help in developing sound and complete arguments, but of course, do not guarantee that such have been produced. Assurance cases are inherently defeasible, which means that there is always the possibility that something has been omitted. When we say that we have "complete" confidence in a claim (i.e., that the claim has Baconian probability $n|n$) we understand that this only reflects what we knew at a particular point in time.

B. Incomplete Defeater Elimination

We speak of "eliminating a defeater" as though a defeater is either eliminated or it is not. What we mean by this is that an argument should state a claim that, *if true*, serves to completely eliminate an associated defeater. Of course, there may be residual doubts about the truth of such a claim, in which case, the same doubts apply to the defeater's elimination. However, as an argument is refined, one eventually develops defeaters that are "obviously" eliminated by associated evidence [13].

For example, consider a claim that the probability of system failure on demand (pfd) is less than 10^{-x}. A counter-example (rebutting defeater) might be "One failure was observed in 10,000 operationally random tests of the system". Evidence eliminating this defeater might be "Log showing no failures observed in 10,000 operationally random tests of the system".

The lack of observed failures might, of course, be meaningless for various reasons, e.g., if the test oracle is unreliable. (The unreliability of a test oracle is an undermining defeater because it attacks the validity of the evidence.) How should we go about demonstrating oracle reliability? We would postulate various reasons that would demonstrate lack of reliability (rebutting defeaters) and try to show that none of these reasons hold. If we can't show that all of them are eliminated, we will be left with residual doubt about oracle reliability, and this doubt propagates to our claim that pfd is less than 10^{-x}.

In short, if it is impossible to eliminate some lowest level defeaters, then the associated doubt leads to incomplete elimination of a higher level defeater, i.e., the elimination of such a defeater remains subject to doubt. The formulation of low level claims that *can* be completely eliminated by appropriate evidence is a goal in developing a convincing argument.

C. Differential Defeater Importance

In any set of defeaters, it is unlikely that they all seem equally important, i.e., intuitively, it can seem that the elimination of one defeater (e.g., the failure of a system that controls the braking in an automobile) may have higher implications for confidence in safety than the elimination of another (e.g., the failure of a system that controls the backup lights in that same automobile.) 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*.

Before continuing, note that in either case we have higher confidence in the claim of system safety than if we were unable to eliminate either of the defeaters. That is, *1|2* (some doubts eliminated) always represents more confidence than *0|2* (no doubts eliminated), and *2|2* (all doubts eliminated) always represents complete confidence.

Although incorporating a notion of relative importance of defeaters into our proposed grading of confidence may seem intuitively desirable, it is not essential to practically use our framework to measure confidence. To understand why, consider the role of hazard analysis in the process of assuring system safety. Hazard analysis helps identify potential safety hazards whose demonstrated mitigation is necessary to establish sufficient confidence that a system is safe. However, not every potential hazard is represented in a safety case because mitigating potential hazards that are conceivable yet extremely unlikely and minimally impactful on system safety would contribute negligible increases to safety. But even though the hazards that are represented in a safety case have different likelihoods and impacts relative to one another, the point of representing them in the safety case is that they *all* must be demonstrably mitigated in order to establish sufficient confidence that the system is safe. As such, assessing the relative importance of hazards is not practically profitable. 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. Consequently, for practical purposes, the notion of relative importance of defeaters is not essential for using our framework to measure confidence.

D. Why Use Baconian Probability?

In enumerative induction, the number of confirming test results provides evidence of the statistical likelihood that future test results will also be confirming. In software, statistical testing [18] is an example of this use of enumerative induction. But given a body of confirming evidence, enumerative induction, by itself, gives us no reason to believe that other tests or analyses yet to be done will also be confirmatory. On the other hand, when a rebutting defeater is eliminated by test or analysis evidence, we have added to our knowledge about why a system works (*cf.* Popper's critical rationalism [19]). In short, with eliminative induction we learn something concrete about why a system works, and with enumerative induction, we at best only learn something statistical about the system (although statistical knowledge can be valuable).

Our approach combines eliminative induction with defeasible reasoning. This combination allows us to articulate not only concrete reasons why a system can fail (the rebutting defeaters) but also concrete reasons why the argument can be defective (the undermining and undercutting defeaters). Uneliminated defeaters give us information about where to focus additional assurance efforts.

In addition to these benefits, the Baconian approach is useful to evaluate a system prior to its operational use; we need to see *what* could go wrong before it *does* go wrong. For example, before a safety-critical system is put into operation, we must be able to reason about possible ways the system could be unsafe and why the system design eliminates or mitigates these possibilities. This is eliminative induction. The assurance case structure helps to structure such reasoning, and our addition of defeasible reasoning concepts provides a framework for explaining why we believe the system is safe, namely, because potential problems (rebutting defeaters) have been identified and eliminated and because possible problems with the argument (undercutting and undermining defeaters) have also been identified and eliminated. Much useful evidence and argumentation can be developed significantly in advance of having an actual operational system. Of course, once a system is operational, we can collect statistics on observed defects (enumerative induction) to predict its future operational reliability and safety.

From a psychological viewpoint, the Baconian approach avoids confirmation bias—the tendency to see all evidence through the perspective of one's chosen theory. In discussing assurance cases, one tends to talk about how evidence "supports a claim", a phrase that can lead one to think that evidence that is consistent with a claim is sufficient to demonstrate the claim's validity. But the question in such thinking always needs to be, "How do we know that situations not covered by the evidence will also be consistent with the claim." The eliminative induction approach gives an answer: to the extent that the evidence eliminates defeaters, we know that the claim cannot be false for all situations covered by these defeaters.

Finally, we use eliminative induction informally all the time as a way of convincing ourselves, or others, of an argument's validity. More formal or structured uses also exist: logical proof by contradiction is one example. Others include Clarke's counter-example guided abstraction refinement for model checking [20] or resilience engineering [21].

None of this is to discount the very real importance of Pascalian approaches to assurance, including Bayesian methods. In fact, we believe they co-exist: one informs the other. If one wants to make Pascalian claims about system reliability, make a Pascalian claim and then elucidate the possibilities that would make you doubt the validity of the claim (as in our *pfd* claim earlier). On the other hand, if you want to determine what the actual operational reliability of a system is, then take repeated samples.

E. Practical Considerations

Although the number of defeaters relevant to an argument seems to be quite large for a real system, the amount of relevant argument and evidence for a real system is inherently quite large. The question is whether the approach of identifying defeaters allows one to develop a more thorough and cost-effective basis for developing confidence in system behavior than current methods — whether this approach leads to more effective and focused assurance efforts. These questions cannot be answered until we have obtained practical experience in applying the approach, but our initial interactions with systems developers have been promising.

V. Summary

The identification of defeaters and how they are eliminated provides a framework for assessing the confidence that one should have in an assurance case. Although the approach is still under development, it provides useful ways of thinking about assurance case deficiencies and the value of eliminating them.

ACKNOWLEDGMENT

We appreciate the helpful comments of the reviewers. They helped us to sharpen our presentation. We'd also like to thank Neil A. Ernst for his valuable contributions the final version of this paper.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

This material has been approved for public release and unlimited distribution. DM-0000186

REFERENCES

[1] T. Kelly, "Arguing Safety -- A Systematic Approach to Safety Case Management," University of York, Department of Computer Science, 1998.

[2] "GSN Community Standard", Origin Consulting, 2011. http://www.goalstructuringnotation.info/documents/GSN_Standard.pdf

[3] ISO/IEC, "15026-2:2011 Systems and software engineering -- Systems and software assurance -- Part 2: Assurance case," 2011.

[4] U.S. Food and Drug Administration. "Guidance for Industry and FDA Staff – Total Product Life Cycle: Infusion Pump – Premarket Notification [510(k)] Submissions".

[5] Ministry of Defence, "Defence Standard – 0056: Safety Management Requirements for Defence Systems – Part 2.", 2007.

[6] R. Hawkins, T. Kelly, J. Knight, and P. Graydon "A New Approach to Creating Clear Safety Arguments", in Advances in Systems Safety: Proceedings of the Eighteenth Safety-Critical Systems Symposium. Southampton, U.K., February 2011. C. Dale and T. Anderson (eds.), Springer-Verlag, 2011.

[7] L. Sun and T. Kelly, "Elaborating the Concept of Evidence in Safety Cases", in Assuring the Safety of Systems: Proceedings of the Twenty-First Safety-Critical Systems Symposium, Bristol, U.K., February 2013. C. Dale and T. Anderson (eds.), Springer-Verlag, 2013.

[8] A. Wassyng, T. Maibaum, M. Lawford, and H. Bherer, "Software Certification: Is There a Case Against Safety Cases," 206-227. Monterey Workshops 2010, Lecture Notes in Computer Science 6662. R. Calinescu and E. Jackson (eds.), Springer-Verlag, 2011.

[9] R. Bloomfield and B. Littlewood, "Multi-Legged Arguments: The Impact of Diversity upon Confidence in Dependability Arguments," 25-34. Proceedings of the International Conference on Dependable Systems and Networks (DSN 2003). San Francisco, CA, June 2003. IEEE, 2003.

[10] R. Bloomfield, B. Littlewood, and D. Wright, "Confidence: Its Role in Dependability Cases for Risk Assessment," 338-346. Proceedings of the International Conference on Dependable Systems and Networks (DSN 2007). Edinburg, U.K., April 2007. IEEE, 2007.

[11] B. Littlewood and D. Wright, "The Use of Multilegged Arguments to Increase Confidence in Safety Claims for Software-Based Systems: A Study Based on a BBN Analysis of an Idealized Example." IEEE Transactions on Software Engineering 33, 5 (May 2007): 347-365.

[12] J. Goodenough, C. Weinstock, and A. Klein, "Toward a Theory of Assurance Case Confidence," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2012

[13] J. Goodenough, C. Weinstock, and A. Klein, "Eliminative Induction: A Basis for Arguing System Confidence," New Ideas and Emerging Results Workshop, International Conference on Software Engineering, in press.

[14] D. Schum, The Evidential Foundations of Probabilistic Reasoning, Northwestern University Press, 2001.

[15] J. Pollock, "Defeasible Reasoning," in Reasoning: Studies of Human Inference and Its Foundations, J. E. Adler and L. J. Rips, Eds., Cambridge University Press, 2008, pp. 451-469.

[16] H. Prakken, "An Abstract Framework for Argumentation with Structured Arguments," Argument & Computation, vol. I, no. 2, pp. 93-124, 2010.

[17] L. J. Cohen, An Introduction to the Philosophy of Induction and Probability, Clarendon, 1989.

[18] J. Duran and S. Ntafos, "An evaluation of random testing," IEEE Transactions on Software Engineering 10, 4 (1984): 438-444.

[19] K. Popper, Conjectures and Refutations: The Growth of Scientific Knowledge, London: Routledge, 1963.

[20] E. Clarke, O. Grumberg, S. Jha, Y. Lu, and H. Veith, "Counter-example guided abstraction refinement", Computer-Aided Verification Conference, pp. 154-169, 2000.

[21] T. Limoncelli, "Case Study: Learning to Embrace Failure", ACM Queue, September, 2012.