

Aggregating Confidence: A Problem and A Solution

Stella Cohen and Rose Gotlieb**
Swarthmore College, Swarthmore, Pennsylvania

C. Michael Holloway
Langley Research Center, Hampton, Virginia

Sarah M. Lehman
Langley Research Center, Hampton, Virginia

** These authors contributed equally to this work.*

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 NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collection 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:

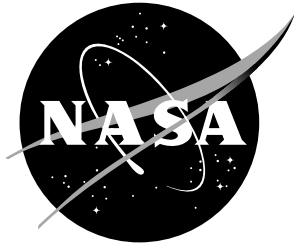
- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing 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



Aggregating Confidence: A Problem and A Solution

Stella Cohen and Rose Gotlieb**
Swarthmore College, Swarthmore, Pennsylvania

C. Michael Holloway
Langley Research Center, Hampton, Virginia

Sarah M. Lehman
Langley Research Center, Hampton, Virginia

** These authors contributed equally to this work.*

National Aeronautics and
Space Administration
Langley Research Center
Hampton, Virginia 23681-2199

November 2023

Acknowledgments

The following people provided ideas, encouragement, and thoughtful criticisms without which the document you are about to read would not exist in its present form: Zamira Daw, Mallory Graydon, and Kimberly Wasson. Thank you!

This work was sponsored in part by the Federal Aviation Administration through Interagency Agreement IA1-30333, Annex 3: “Using the Overarching Properties in Novel Examples (OPiNE)”, with Srinu Mandalapu as the technical contact.

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

This paper has two main purposes: to articulate a concern about assessing ARGUMENTS based on confidence; and to provide an assessment procedure which takes this concern into account. The concern is that confidence is a tricky notion for non-deductive ARGUMENTS: doubts inherent in REASONING coalesces with doubts inherent in PREMISES. And, within a COMPOUND ARGUMENT, this may happen many times over, thus accumulating in unpredictable ways. We explain two procedures we have developed to help address the concern in a methodical way.

1 Introduction

The purpose of this paper is to indicate and explore a concern about a particular kind of ARGUMENT, the type that is used to demonstrate whether a subject possesses the Overarching Properties (OPs). The intended audience of this paper is twofold: primarily, individuals working on developing assessment tools for such ARGUMENTS, and secondarily, those who want to use the OPs. The OPs are intended to provide a framework from which confidence in the safety of aviation systems may be justified using non-deductive ARGUMENTS¹, called Overarching Properties-Related Arguments (OPRAs).

OPRAs are meant to *justify confidence*. To ask what confidence is, or what it means to have it justifiably, is to open a can of worms. Confidence is not, for instance, simply a subjective or objective feature of claims; its relationship with probability is nonzero but not one-to-one; and it is, by nature, vague. Luckily for us, we do not have to open the “what is confidence” can in this paper; rather, we will rely on the intuitive sense of the word. We do, however, have to open a similar can: how does confidence *work*?

Elucidating how confidence works is crucial for OPRAs for the simple reason that confidence is conceptually central to OPRAs. However, the question’s importance runs deeper than that. In this paper, we identify a concern about how confidence works in OPRAs when it is not given sufficient attention. Namely, the concern is that in large-scale ARGUMENTS,² it is very difficult to track how the confidence in the premises is transmitted into confidence in the conclusion. It seems as though confidence can often “leak” throughout an ARGUMENT and leave you less confident in your conclusion than you are in your premises. This concern about leaking confidence, which we will present in detail in Section 3, is applicable to all large-scale ARGUMENTS whose premises are open to doubt. However, as this paper is particularly concerned with OPRAs, examples and discussion will follow their form rather than trying to address all types of ARGUMENTS.

To ignore this concern would endanger the ARGUMENT assessment—in the worst-case scenario, to ignore this concern might lead assessors to accept ARGUMENTS as

¹For clarity on this term, see Section 2.4.

²By “large-scale” ARGUMENTS, we mean ones with many premises and, in particular, with many premises that are defended by their own ARGUMENTS.

COGENT which are, in reality, not COGENT at all. Addressing this concern is the impetus for our paper.

In what follows, we will begin with a brief overview of key components of OPRAs and related documents which detail how to construct and assess them (Section 2). We will then describe in detail our concern about confidence (Section 3). In Section 4, we will demonstrate how this concern can cause problems for the ARGUMENT assessment tool iTest. In Section 5, we will provide a possible procedure for assessing ARGUMENTS which takes this concern about confidence into account. In Section 6, we will detail possible directions readers may want to explore in response to this paper. Our last section will offer a brief summary of what we have concluded.

2 Background

For readers unfamiliar with the OPs or with OPRAs, there are several documents worth reading before continuing with this paper:

1. *A Primer on Argument* [1] establishes an accessible, consistent lexicon for talking about and working with structured ARGUMENTS.
2. *The Friendly Argument Notation (FAN)* [2] establishes an accessible, text-based, platform-agnostic notation for writing structured ARGUMENTS.
3. *A Primer on Argument Assessment* [3] provides guidelines for a way (but not the only way) to assess ARGUMENTS.
4. *Understanding the Overarching Properties* [4] introduces the properties, their origins, definitions, and relevance.
5. *An Introduction to Constructing and Assessing Overarching Properties-Related Arguments (OPRAs)* [5] introduces construction and assessment of ARGUMENTS in relation to the Overarching Properties.³

Ideally, readers of this paper will be fluent in the documents above. That said, as time is a precious resource, we will here provide a barebones overview of the first three documents, which are most pertinent for comprehension of our paper. This overview best serves as a refresher, rather than an introduction to these concepts, but may also suffice as explanation for some readers less familiar with the work. However, if you belong to the latter category and at some point throughout this paper become confused, or curious about more context, it will likely be helpful to return to the documents themselves.

As we introduce a possible concern and procedure in this paper, we will draw upon the above documents. However, note that, in some places, our claims, precepts, and definitions may differ from those presented in the above. We will make these differences sufficiently clear when they arise.

³The summaries of these papers used below are based on, and in appropriate cases copied directly, from this paper.

2.1 *A Primer on Argument*

A Primer on Argument identified a collection of primitives, precepts, and practical considerations important for discussing, creating, and assessing ARGUMENTS. The most salient part for this paper is understanding the primitives (or if you prefer “terms”). We present the definitions without discussion.

- An **ARGUMENT** is an attempt to convince others to BELIEVE a CONCLUSION through REASONING and one or more PREMISES.
- To **BELIEVE** is to accept as true.
- The **CONCLUSION** is the statement you want your audience to BELIEVE.
- A **PREMISE** is a statement you think your audience BELIEVES.
- **REASONING** is the statement of why you think the PREMISES should cause your audience to BELIEVE your CONCLUSION.
- A **BINDING** is an association between a term used in an ARGUMENT and the real-world information to which that term refers.
- A **DEFEATER** is a statement that may cause your audience to not BELIEVE your CONCLUSION.
- An **ATOMIC ARGUMENT** consists of a single CONCLUSION together with its immediate REASONING, PREMISES, BINDINGS (if present), and DEFEATERS (if present).
- A **COMPOUND ARGUMENT** is an ARGUMENT consisting of more than one ATOMIC ARGUMENT.
- An ARGUMENT is called **COGENT** if it rationally justifies BELIEVING its CONCLUSION to the required standard of confidence.

These primitives express the goal of an ARGUMENT (BELIEF) and encompass the necessary elements of an ARGUMENT: CONCLUSION, REASONING, and PREMISE(s). A CONCLUSION alone is not an ARGUMENT. REASONING alone is not an ARGUMENT. A PREMISE or several PREMISES is not an ARGUMENT. Several PREMISES and some REASONING do not an ARGUMENT make, and even a CONCLUSION with REASONING and one or more PREMISES is not an ARGUMENT, unless it is being offered for the purpose of trying to convince someone to BELIEVE the CONCLUSION.

All of these primitives are accepted and used in this manner in the same way that they are presented in *A Primer on Argument*.⁴ In addition, when they arise throughout the rest of the paper, they will be in SMALL CAPS as a reminder of their specific meanings.

⁴This summary of *A Primer on Argument* primarily comes from Holloway and Wasson’s “A Primer on Argument Assessment” [3].

2.2 The Friendly Argument Notation (FAN)

The Friendly Argument Notation (FAN) was developed to facilitate the creation, understanding, and assessment of structured ARGUMENTS, with a particular (but not exclusive) focus on ARGUMENTS about safety-critical systems. FAN intentionally corresponds closely to traditional ARGUMENT concepts. It is designed to be efficient to write and read, with sufficient semantic and syntactic rigor to capture the necessary distinctions. Consider the famous, ancient ARGUMENT for Socrates' mortality. Given the two PREMISES, "All men are mortal" and "Socrates is a man", we conclude that "Socrates is mortal". In FAN, this ARGUMENT might look like this:

Believing
 Socrates is mortal
is justified by applying
 the AAA-1 syllogism
to these premises
 All men are mortal
 Socrates is a man

Throughout this paper, we will put some example ARGUMENTS in FAN for ease of reading and to give readers practice in assessing its particular syntax.⁵ However, at times, also for ease of reading, we will write out the ARGUMENTS more casually in the midst of a paragraph, simply relaying our PREMISES and our desired CONCLUSION. In many of these situations, we will not overtly state our REASONING, especially when the only REASONING is "conjunction". For example, conjoining the PREMISES "I'm tired, it's 11pm", and "I have nothing else to do tonight" is a fair way to reason towards the CONCLUSION "I should go to bed". However, when articulating this ARGUMENT casually, it may not be necessary to state that I am using "conjunction". In sum, do not think that we are genuinely omitting REASONING from our casual articulations of ARGUMENTS. REASONING is always present.

2.3 A Primer on Argument Assessment

One way to assess an ARGUMENT is presented in *A Primer on Argument Assessment*, namely, iTest. It focuses on using questions to guide the assessment of whether (a) the form of an ARGUMENT is proper, and (b) the content is sufficient to justify believing the CONCLUSION. iTest is used for COMPOUND ARGUMENT, but it iterates a second procedure, SPRY, which assesses ATOMIC ARGUMENTS. We will begin by overviewing SPRY.

⁵This summary of FAN largely comes from Holloway and Wasson's "A Primer on Argument Assessment" [3].

SPRY is made up of four basic questions, where an affirmative response means that you can move to the next question, and all affirmative responses means that the ARGUMENT is COGENT. The four questions are:

1. Is the **Syntax** proper?
2. Are the **PREMISES** acceptable?
3. Is the **REASONING** acceptable?
4. Is saying "Yes" to the CONCLUSION justified?

One way to answer the "S" question is to answer these additional questions:

1. Is there a single CONCLUSION that is stated in the form of a proposition?
2. Is there a statement of REASONING?
3. Is there at least one PREMISE?
4. Is each PREMISE stated in the form of a proposition?
5. If there are DEFEATERS, is each stated in the form of a proposition?
6. Does a BINDING exist for each term or phrase used in the CONCLUSION, REASONING, PREMISES, and (if any) DEFEATERS that do not have a well-known, unambiguous definition?
7. Does a proper BINDING exist for each reference to an external artifact?

One way to answer the "P" question is to answer these additional questions:

1. Is each PREMISE BELIEVABLE? To be BELIEVABLE, it must fall into one of these categories:
 - (a) Expresses a proposition that is 'universally' accepted as true
 - (b) Expresses a proposition that is accepted as true within the relevant domain
 - (c) Is supported by an ARGUMENT (provisionally presumed to be COGENT)
 - (d) Is supported by external artifacts that are accepted within the domain as being sufficient to establish its truth
 - (e) Will fall into category (c) or (d) at a later stage of ARGUMENT development
 - (f) Is an assumption accepted by all stakeholders and clearly identified as such

2. Is each PREMISE relevant to the CONCLUSION?

One way to answer the “R” question is to answer these additional questions:

- a. Is the REASONING relevant?
- b. Is the REASONING consistent with current knowledge?

One way to answer the “Y” question is to answer these additional questions:

- a. Is the required level of confidence known?
- b. Does the ARGUMENT engender the required level of confidence?

One way, but not the only way, to assess the COGENCY of a COMPOUND ARGUMENT is to apply iTest. Applying iTest involves completing the following steps:

- I. Isolate one ATOMIC ARGUMENT.
- II. Interrogate it as an ATOMIC ARGUMENT using the SPRY questions.
- III. Iterate until all ATOMIC ARGUMENTS have been assessed.
- IV. Integrate the individual assessments by answering these questions:
 - a. Are all ATOMIC ARGUMENTS assessed as COGENT?
 - b. Does every path through the COMPOUND ARGUMENT terminate in BELIEVED PREMISES?

One way, but not the only way, to propagate the effect of non-COGENT ATOMIC ARGUMENT ON a COMPOUND ARGUMENT is to follow these five steps:

1. For each non-COGENT ATOMIC ARGUMENT, list its CONCLUSION.
2. For each listed CONCLUSION, find all ATOMIC ARGUMENTS for which it serves as a PREMISE. Note that each such PREMISE will have been identified as falling into BELIEVABILITY category (c).
3. Change the PREMISE’s BELIEVABILITY from category (c) to “No.”
4. Reassess the COGENCY of the ARGUMENT in which the PREMISE appears.
5. Repeat until no new non-COGENT ARGUMENTS are revealed.

We will provide more commentary on SPRY and iTest in Section 4, so, if this process is not sufficiently comprehensible to you, please read through the pertinent documents before beginning that section.⁶

⁶This summary of SPRY and iTest comes from the Appendix of Holloway and Wasson’s “An Introduction to Constructing and Assessing Overarching Properties-Related Arguments (OPRAs)” [5].

2.4 Confidence

The primitives of ARGUMENTS and FAN describe ARGUMENTS, and provide a way to express them. However, this paper is concerned with what makes an ARGUMENT *good*. Thus, some clarity on the goal of ARGUMENTS is called for.

The goals of ARGUMENTS differ depending on whether an ARGUMENT is deductive or non-deductive. What is the primary difference between a deductive and a non-deductive ARGUMENT? The former is an ARGUMENT where true PREMISES are combined with REASONING which absolutely guarantees that your CONCLUSION is true. Non-deductive ARGUMENTS, on the other hand, are any ARGUMENTS where that is *not* the case. One example of a non-deductive ARGUMENT is one which uses evidence to support claims. If I have a *preponderance of evidence* in support of the claim that a particular window was broken by the impact of a particular rock, this claim might still be challenged by a subsequent piece of evidence—i.e., a person confessing to having punched it.⁷ Non-deductive ARGUMENTS can be very strong, but new evidence can always come to light: they cannot be certain.

When an ARGUMENT is “good”, it is called COGENT. When making a deductive ARGUMENT, COGENCY requires that the PREMISES be *true* and the REASONING be *valid*, for only true PREMISES and valid REASONING guarantee a true CONCLUSION. But in the context of OPRAs, COGENCY is all about *confidence*: for an ARGUMENT to be COGENT, it must rationally justify believing its CONCLUSION to the required standard of confidence. We have to define COGENCY this way, both because most OPRAs deal with non-deductive REASONING, and because makers and assessors of OPRAs are not omniscient: absolute truth, in this context, would be too high a standard.

What does the requirement that ARGUMENTS engender *confidence* in their CONCLUSIONS entail? Confidence is a difficult concept to define, and we believe that the intuitive understanding is strong enough to justify not providing a specific definition of it. Instead, we will briefly provide two examples of confidence standards, both from the American legal system.⁸

The required level of confidence for a judge or jury to convict a criminal defendant is (with some slight variations in wording depending on the state) “beyond a reasonable doubt”. In contrast, to find for the plaintiff in a civil action, the judge or jury need only reach a “more likely than not” level of confidence. An ARGUMENT that is COGENT for a civil action may well not be COGENT for a conviction in a criminal trial. This goes to show that the purpose and context of the ARGUMENT determines what standard of confidence should be used. What counts as “confident” will depend heavily on your situation—which, as shown in the case of the legal system, is a familiar and reasonable way of assessing non-deductive ARGUMENTS. However, using *confidence* to determine COGENCY instead of *truth* will inevitably lead to some difficulties. These we will turn to now.

⁷Our use of the term “evidence” and this example are highly indebted to “Towards a Coherent View of Evidence in Safety Assurance” by McCardel, Holloway, Wasson, McDonnell, Graydon, Peña, and Lehman [6].

⁸This example is taken from [3].

3 The Concern of Leaking Confidence

The concern we want to highlight is that throughout an ARGUMENT, lack of confidence can aggregate in ways that are difficult to predict. So, when a CONCLUSION depends upon many PREMISES, each of which are not absolutely true, all of those small deficits in confidence can accumulate, thereby making one's confidence in the CONCLUSION lower than one's confidence in any one of the PREMISES.

To see how this might occur, consider the following ARGUMENT: I am fairly confident that I am going to go for a walk today. I am fairly confident that, if I go on a walk, I will go on a walk in the park near my house. I am fairly confident that, if I go on a walk in the park near my house, I will do the two-mile loop that takes me through the woods. I am fairly confident that, if I do this two-mile loop, I will also have enough energy to walk home instead of taking the bus. I am fairly confident that, if I walk home, I will take the route that passes by the ice cream store. And I am fairly confident that, if I pass the ice cream store, I will end up going in and getting ice cream.

This stream of conditionals can be understood as a COMPOUND ARGUMENT wherein each hypothetical is a single PREMISE for a CONCLUSION, which is then taken to be a PREMISE for the next level of the ARGUMENT. So, the CONCLUSION "I will walk in the park" is concluded from the PREMISE "I will go on a walk" and the REASONING "If I go on a walk, I will do so in the park near my house"; after being thus concluded, it is used as a PREMISE for the subsequently desired CONCLUSION, "I will do the two-mile loop". For a visual:

Believing

I will get ice cream

is justified by applying

Modus Ponens

to these premises

If I pass the ice cream store, I will get ice cream {i}

I will pass the ice cream store {ii}

Believing

I will pass the ice cream store

is justified by applying

Modus Ponens

to these premises

If I walk home, I will take the route that passes the ice cream store {i}

I will walk home {ii}

Believing

I will walk home

is justified by applying

Modus Ponens

to these premises

If I do the two-mile loop, I will have the energy to walk home {i}

I will do the two-mile loop {ii}

The problem is that each time I move up a step in the ARGUMENT, I lose confidence. The PREMISE “I will go on a walk” is already only fairly confident. The REASONING “If I go on a walk in the park, I will go on a walk in the park near my house”, too, only engenders a fair amount of confidence to begin with—that is, even if I were absolutely confident that I was going for a walk, I would *still* only be fairly confident about doing so in the park. REASONING which engenders only fair confidence, in combination with an only fairly confident PREMISE, would seem to engender a *less* than fairly confident CONCLUSION—which, then, we are using to defend our *next* PREMISE!⁹ By the time I get to wondering about the ice cream store, I’ll have very little confidence at all.

This example illustrates how, when you move through an non-deductive ARGUMENT made up of multiple smaller ARGUMENTS, lack of confidence can aggregate; this means that, however confident you are in your individual cases of REASONING and your individual PREMISES, you may be far less confident in your ultimate CON-

⁹Why, you might ask, would anyone use REASONING which engenders only fair confidence? Well, in truth, no non-deductive ARGUMENT can guarantee the truth of a CONCLUSION to perfect confidence. Any non-deductive ARGUMENT will contain seeds of doubt in the REASONING itself—which does not mean that the REASONING is *flawed*, simply that it does not guarantee the CONCLUSION. The real world is messier than dedicated formalists want it to be.

CLUSION.¹⁰ Due to leaky confidence, *your CONCLUSION can often be weaker than any of its supporting elements.*

This example seems to evince a pattern wherein the more steps you have in your ARGUMENT, the more confidence leaks, and thus the greater your lack of confidence is in your CONCLUSION. If you tend towards using mathematics, you might be tempted to assign a confidence score to each PREMISE and write a function which intakes individual confidence scores and outputs a final confidence score in your CONCLUSION. In reality, however, lack of confidence aggregates in ways that are more complicated than the above example, and that thus will make attempts at this type of systematization much more difficult.

To see this difficulty a bit more clearly, consider the following ARGUMENT. I want to defend the CONCLUSION that I will get ice cream today. I have the following PREMISES, each of which I have defended with their own ATOMIC ARGUMENT, to particular degrees of confidence:

1. I am “very confident” that I really want ice cream.
2. I am “absolutely confident” that there is an ice cream place at the end of my walk.
3. I am “very confident” that I am taking a walk today.
4. I am “somewhat confident” that on my walk, I will meet my friend Harry.
5. I am “very confident” that Harry loves ice cream and would ask me to go with him to the store.
6. I am “not so confident” that the store carries my favorite flavor today.
7. I am “not so confident” that waffle cones are half-off today.

By looking at this ARGUMENT, it seems reasonable to say that I am very confident that today, I will be getting some ice cream. Reasonable doubts abound, of course, but considering all of these levels of confidence—I really want ice cream, I’ll pass the store, and it’s not unlikely that Harry will drag me there anyways—the future is looking sweet.

In simple mathematical models of determining the confidence in the CONCLUSION based on the confidence scores of the PREMISES, the lower confidences in PREMISES (6), (7), and perhaps (4) and thus (5) might make you seem *less* confident than you really ought to be. This example goes to show that in non-deductive ARGUMENTS, certain PREMISES might be more relevant to your CONCLUSION than others—perhaps on their own, perhaps in combination with other PREMISES. The way that the confidence level of each PREMISE bears upon the confidence level of the CONCLUSION

¹⁰Though the example ARGUMENT we just used only includes if-then statements as PREMISES, this concern can arise with many other types of REASONING; this example is not just a fringe case.

meaningfully depends upon the way in which the PREMISE relates to the CONCLUSION in the first place. Similarly, the *content* of the PREMISE—that is, what it says about the truth of the CONCLUSION—is important to keep in mind, for “I am very confident that I really want ice cream” seems to have more potential to guarantee the desired CONCLUSION than, say, “I am very confident that waffle cones are half-off today”. Additionally, the *context* of the PREMISE—that is, its relationship with other PREMISES—is also important, for my absolute confidence in the location of the ice cream store (2) bolsters my high confidence in my walk leading to ice cream (3).¹¹ We cannot create a function that simply takes in confidence scores of PREMISES and provides the confidence in the CONCLUSION; *Confidence aggregation depends on content and context.*¹²

In this section, we have shown that lack of confidence aggregates, and in a way that is difficult to predict. This is a concern for OPRAs, which are large-scale COMPOUND ARGUMENTS which use non-deductive REASONING. In what follows, we will bring this concern to bear on current methods of ARGUMENT assessment, and provide a tool whereby one can keep close track of aggregating confidence throughout ARGUMENT assessment.

4 Leaking Confidence and iTest

As seen in the previous section, keeping track of confidence levels throughout the assessment of a COMPOUND ARGUMENT is of utmost importance; confidence often “leaks” out of an ARGUMENT, and it does so in ways that are not easily predictable. Any method you use of assessing ARGUMENTS should thus account for this difficulty. In this section, we will explore one current method of assessing OPRAs—iTest, which we have glossed in Section 2—to demonstrate how this difficulty can arise.

When evaluating a COMPOUND ARGUMENT using iTest, an assessor performs small SPRY evaluations on each ATOMIC ARGUMENT, in no particular order, iterating until all ARGUMENTS have been deemed COGENT. If all ATOMIC ARGUMENTS are COGENT and a few structural conditions on the COMPOUND ARGUMENT are satisfied, then the overall ARGUMENT is considered COGENT. However, the final question in SPRY (“Is saying Yes justified to the required level of confidence?”) can become problematic. It is impossible to answer this question unless you decide *what* the required level of confidence is for this particular ATOMIC ARGUMENT. iTest is flexible here, allowing assessors to decide what standards would be appropriate.

Though this flexibility means that assessors *can* account for leaking confidence if they manage to choose all the right standards of confidence, it does not give assessors much guidance for *how* to choose them. As you cannot complete the final step of SPRY without a specified standard of confidence, the rest of this section will

¹¹In Section 5, we will describe in more detail what it means for a PREMISE to depend on its context. For now, just note that the relationships between PREMISES can be important.

¹²This is not to say that confidence aggregation can *never* be mathematically formalized. Our claim is that it is difficult, not necessarily impossible. In [Section 6.2](#), we provide further discussion on the difficulties of quantifying confidence.

be devoted to considering two different intuitive answers to assist in determining the required standard of confidence. Ideally, we would find a way to specify required standards of confidence so that iTest would accept ARGUMENTS if and only if they are COGENT—that it would neither approve of non-COGENT ARGUMENTS nor keep COGENT ARGUMENTS from being so declared.

One possibility would be to hold each ARGUMENT to the same standard of confidence. That is, if you want your final CONCLUSION to be “very likely”, you would hold every atomic CONCLUSION to the standard of “very likely”. However, this method makes iTest too weak in that it allows iTest to accept ARGUMENTS that do not actually engender as much confidence as desired. Consider the following example: Say that Ashley has just handed me the following ARGUMENT that Ms. Scarlet killed Mr. Boddy, and I am now going to apply iTest to determine if it justifies “fair confidence”. For the sake of this example, some ARGUMENTS will be written in FAN but with placeholders for the REASONING and PREMISES rather than real content.

Believing

Ms. Scarlet killed Mr. Boddy

is justified by applying

Conjunction

to these premises

Mr. Boddy did not die of natural causes {i}

Ms. Scarlet had a motive {ii}

Ms. Scarlet was in the house when Mr. Boddy died {iii}

No one else was in the house when Mr. Boddy died {iv}

Believing

Mr. Boddy did not die of natural causes {i}

is justified by applying

⟨ Reasoning A ⟩

to these premises

⟨ Premise 1 ⟩

⟨ Premise 2 ⟩

Believing

Ms. Scarlet had a motive {ii}

is justified by applying

⟨ Reasoning B ⟩

to these premises

⟨ Premise 3 ⟩

⟨ Premise 4 ⟩

Believing

Ms. Scarlet was in the house when Mr. Boddy died {iii}

is justified by applying

⟨ Reasoning C ⟩

to these premises

⟨ Premise 5 ⟩

⟨ Premise 6 ⟩

Believing

No one else was in the house when Mr. Boddy died {iv}

is justified by applying

⟨ Reasoning D ⟩

to these premises

⟨ Premise 7 ⟩

⟨ Premise 8 ⟩

As the assessor, I decide to begin by considering the ARGUMENT for the final CONCLUSION, the ARGUMENT that Ms. Scarlet killed Mr. Boddy. I apply SPRY to the

ARGUMENT and note that it is syntactically proper. Its PREMISES all fall into category (c) of BELIEVABILITY (as described for SPRY), i.e., they are the CONCLUSIONS of their own local ARGUMENTS. Its PREMISES are also relevant, so they are acceptable. The REASONING is acceptable. Further, though this part is somewhat up to the assessor's judgment, I would claim that this justifies saying "yes" to the CONCLUSION to the level of "fair confidence". I certainly would not bet my life on it, but this ARGUMENT seems to rationally justify "fair confidence". Therefore, this ATOMIC ARGUMENT passes SPRY.

For the sake of this example, let us assume REASONING A-D and PREMISES 1-8 make all the other ATOMIC ARGUMENTS COGENT to the level of "fair confidence" as well. Then, because every ATOMIC ARGUMENT has been assessed as COGENT, and every path through the COMPOUND ARGUMENT terminates in BELIEVED PREMISES, iTest would direct me to mark this ARGUMENT as COGENT.

However, this ARGUMENT as a whole should *not* be marked COGENT: while the ARGUMENTS for PREMISES (i-iv) engender fair confidence, the complete argument seems to warrant less confidence in its CONCLUSION. If I am only *fairly confident* that Mr. Boddy did not die of natural causes, and *fairly confident* that Ms. Scarlet had a motive, and *fairly confident* that Ms. Scarlet was in the house when Mr. Boddy died, and *fairly confident* that no one else was in the house when Mr. Boddy died, I should not be confident at all that Ms. Scarlet is the murderer—only, perhaps, suspicious.

The problem with this method is that it obscures leaking confidence. Because we were "fairly confident" in our four PREMISES, iTest approved of them. But then, when we look at the final CONCLUSION, we simply check if those PREMISES, *regardless of their individual confidence levels*, would lead to the desired CONCLUSION with the associated confidence level. If I were *absolutely confident* in PREMISES (i-iv), then their conjunction would engender fair confidence that Ms. Scarlet is the murderer—but I am *not* absolutely confident in them.

This counterexample is representative, not of a fringe case, but of a substantive problem. It reflects what we demonstrated in Section 3: "*fairly confident*" PREMISES in combination with "*fairly confident*" REASONING do not ensure a "*fairly confident*" CONCLUSION.

In response to seeing this method, which approves of non-COGENT ARGUMENTS, we might instead look to a method which aims more pointedly at keeping non-COGENT ARGUMENTS from being approved. That is, in a second method, the assessor could decide that all ATOMIC ARGUMENTS (besides the final CONCLUSION) must be something like "absolutely confident"—i.e., as confident as I am that I have two hands or that the sun will rise tomorrow. Facing the above specification, which let bad ARGUMENTS in, this standard is quite appealing, as it seems a good way to keep them out; if at every stage of assessment, you reject anything in which you are not absolutely confident, it would seem there is no lack of confidence to aggregate. Then, when you check the final CONCLUSION, you can hold it to whichever standard you would like and simply look if the PREMISES and the REASONING are enough to engender that level of confidence. You do not have to be worried about any hidden lack of confidence in the PREMISES.

However, this standard seems to lead us to an opposite problem: it makes iTest

reject ARGUMENTS that actually are COGENT. Consider the following example: I want to be “very confident” that Ms. Scarlet murdered Mr. Boddy. But now, my PREMISES are (a) Ms. Scarlet was found holding the murder weapon, (b) Ms. Scarlet’s journal includes detailed plans about how to kill Mr. Boddy and (c) Ms. Scarlet was recorded in the act of the crime on fifteen security cameras. If I am “very confident” but not “absolutely confident” about these PREMISES, I should still be at least “very confident” that Ms. Scarlet is the murderer. Since these are better reasons to BELIEVE the CONCLUSION than the PREMISES (i-iv) from the previous example,¹³ it is possible to be less confident in them individually. If we hold every PREMISE to a level of “absolutely confident”, we are no longer able to use *any* PREMISE that relies on a less than “absolutely confident” ARGUMENT; I would not be able to use the “very confident” PREMISE that Ms. Scarlet was caught in the act by fifteen security cameras *at all*, even though it is clearly information that should help me justifiably conclude she is the murderer. Even if I have some doubts about these PREMISES, they should be able to bolster quite a bit of confidence in my CONCLUSION. But by holding each local ARGUMENT to the standard of absolute confidence, (a, b, c) would be rejected as PREMISES, and the ARGUMENT would fail.

This counterexample, too, is representative of a substantive problem. When making ARGUMENTS in the real world, we often will not have any ATOMIC ARGUMENTS strong enough to make us “absolutely confident”. There are simply too many doubts that one could raise. However, that does not mean that we should not be able to create COGENT ARGUMENTS with the evidence we have. Rather, we want a system that allows us to use less than “absolutely confident” PREMISES which also keeps a close eye on the confidence gap and how it aggregates.

We noted in the previous section that confidence aggregation depends on a PREMISE’s *content* and *context*. The problem with both of the approaches detailed in this section is that, by setting a standard of confidence for local ARGUMENTS across the board, you are unable to consider each PREMISE’s individual content and context. To adequately set a confidence standard for each local ARGUMENT, it would seem necessary to be able to understand its position in and relationship to the broader COMPOUND ARGUMENT. But, because in iTest an assessor can iterate SPRY in any order, PREMISES are removed from their context from the outset. This makes it difficult if not impossible to provide adequate guidance on how to specify required standards of confidence for local ARGUMENTS within a COMPOUND ARGUMENT.

5 SLOW and STAIID

We have demonstrated the need for a procedure which allows an assessor to maintain flexibility when setting the confidence standards for local ARGUMENTS. We have seen from both of the examples above that, when a standard of confidence is set for all the local ARGUMENTS, the standard tends to be either too weak or too strong: it

¹³By better reason, we mean something like more directly relevant, more directly guaranteeing. Say we have one ARGUMENT that Ms. Scarlet is the murderer with the PREMISE “Ms. Scarlet had a motive” and a second ARGUMENT that is identical, but with the PREMISE “Ms. Scarlet was found holding the murder weapon” instead. The second ARGUMENT is almost guaranteed to be stronger.

either approves of non-COGEN arguments or rejects COGEN ones. We have also seen that the content and the context of each PREMISE matter to what its standard of confidence should be. Therefore, what is required is a procedure which allows the assessor to maintain flexibility in setting confidence standards when approaching *each individual local ARGUMENT*, rather than *en masse*.

Not only do different PREMISES have different degrees of importance in justifying CONCLUSIONS, but, importantly, these degrees are not set in stone. What does this mean? In constructing a non-deductive ARGUMENT, there are many different ways the argumentative weight can be balanced between PREMISES. Imagine an ARGUMENT with a set of PREMISES (a-f) and CONCLUSION (g). Perhaps if (b) is certain, then (g) is guaranteed; or, if (b) is only fairly confident, but taken in conjunction with high confidence in (d) and (f), the CONCLUSION holds; or, in a different scenario, if we know nothing about (b), but have high confidence in (a) and (e) and moderate confidence in (c), then we can still conclude (g). An assessor could not, just by looking at the PREMISES, set each PREMISE's necessary confidence level, for *they have no necessary confidence levels*: there are many different ways the confidence could be balanced.¹⁴

Thus, in what follows, we will propose a procedure which does not *set* standards of confidence for each local ARGUMENT. Rather, it *evaluates* the confidence level that each local ARGUMENT guarantees, and then, considering those varying levels, determines how much confidence we can have in our final CONCLUSION. Thus, the assessor can retain flexibility in assessment; they can call ARGUMENTS COGEN with different levels of confidence engendered by different local ARGUMENTS, and they do not have to try every possible combination of confidence levels, which would be unwieldy and likely impossible.

The procedure is made up of two assessment tools: one which is used upon each local ARGUMENT (SLOW), and one which guides the assessor through the overall COMPOUND ARGUMENT (STAID). We will first describe SLOW, followed by STAID. Preliminarily, however, we will introduce some vocabulary.

5.1 Key Terms

There are six important terms we will be using throughout the following sections, which we will define here. They are **terminal PREMISES**, **defended PREMISES**, **assessed PREMISES**, **unassessed PREMISES**, **locked ARGUMENTS**, and **unlocked ARGUMENTS**.

- **Terminal PREMISES:** A terminal PREMISE is a PREMISE which is not the CONCLUSION of any local ARGUMENT. It may seem that everything should be defended,

¹⁴It is certainly possible to assess an ARGUMENT by imposing *one way* you want it to be balanced. You could look at the ARGUMENT, determine that having PREMISE (b) be fairly confident and PREMISES (d) and (f) be highly confident means your CONCLUSION is sufficiently confident. However, with this strategy, you would have to reject ARGUMENTS if one PREMISE is not as confident as you would like. That rejection would not be very meaningful because there are so many other ways the ARGUMENT could be balanced. The ARGUMENT could still be sufficient in its current state.

but we must remember that we should only go as deep as necessary. Much as we would like to argue forever, the buck must stop somewhere.

- **Defended PREMISES:** A defended PREMISE is a PREMISE which *is* the CONCLUSION of its own local ARGUMENT. Note: just because a defended PREMISE is *defended* does not mean it is defended *well*. A PREMISE whose justification is a local ARGUMENT is called a *defended PREMISE* whether or not the ARGUMENT turns out to engender much confidence. All PREMISES are either *terminal PREMISES* or *defended PREMISES*.
- **Assessed PREMISES:** An assessed PREMISE is a PREMISE (terminal or defended) whose confidence level has been determined by the assessor.
- **Unassessed PREMISES:** An unassessed PREMISE is a PREMISE (terminal or defended) whose confidence level has *not* been determined by the assessor. Before any assessment has begun, all PREMISES will be **unassessed PREMISES**. By the time an ARGUMENT is being determined COGENT or not, all PREMISES should be **assessed PREMISES**. Needless to say, no PREMISE can be both assessed and unassessed at the same time.
- **Unlocked ARGUMENTS:** An unlocked ARGUMENT is a local ARGUMENT which contains only **assessed PREMISES**. We can only evaluate ARGUMENTS which are unlocked.
- **Locked ARGUMENTS:** A locked ARGUMENT is any local ARGUMENT which contains at least one **unassessed PREMISE**. We cannot yet assess locked ARGUMENTS. So, if an ARGUMENT is locked, it is (for now) inaccessible: do not go in. Once an ARGUMENT's PREMISES have been assessed, it becomes **unlocked**.

5.2 5.2 SLOW

SLOW is the four-step assessment tool for local ARGUMENTS. The goal of SLOW is to determine how much confidence the local ARGUMENT engenders in the truth of its CONCLUSION. If you are familiar with SPRY, SLOW has many similarities—but a few key differences as well. SLOW is comprised of four basic components (which we will expound upon in a moment):

1. Is the **Syntax** proper?
2. Is the REASONING **Legitimate**?
3. **Organize** dependence relations.
4. **Weigh** your CONCLUSION.

5.2.1 Is the Syntax Proper?

You are in luck—checking the **Syntax** in SLOW begins exactly the same as checking the **Syntax** in SPRY, a process in which you already (hopefully) have some expertise. As a quick refresher, the list of syntax questions is comprised of the following:

- a. Is there a single CONCLUSION that is stated in the form of a proposition?
- b. Is there a statement of REASONING?
- c. Is there at least one PREMISE?
- d. Is each PREMISE stated in the form of a proposition?
- e. If there are DEFEATERS, is each stated in the form of a proposition?
- f. Does a BINDING exist for each term or phrase used in the CONCLUSION, REASONING, PREMISES, and (if any) DEFEATERS that does not have a well-known, unambiguous definition?
- g. Does a proper BINDING exist for each reference to an external artifact?

SLOW simply adds one new question about the **Syntax**:

- h. Is your ARGUMENT **unlocked**?

You must make sure that your ARGUMENT is **unlocked** and thus that the confidence levels for each PREMISE, whether terminal or defended, is known. This knowledge will be crucial in the steps that follow.

5.2.2 Is the REASONING Legitimate?

This section of SLOW is analogous to the “REASONING” section of SPRY; it is the section in which you check for flaws in the ARGUMENT’S REASONING.¹⁵ There are two questions to ask in this section.

1. Is the REASONING *relevant*?
2. Is the REASONING *consistent with current knowledge*?

As noted in SPRY, it is difficult to give objective guidelines regarding what makes REASONING relevant in this type of ARGUMENT. Any relevant REASONING should, of course, help you get from your PREMISES to your CONCLUSION—if the REASONING fails to do this, it is certainly irrelevant in the most general sense of the word. However,

¹⁵As such, much of the following section is reworked material from [3] p. 13-14

you should also watch out for REASONING which builds a false bridge from the instance, the fallacy of division. We want to conclude that *Joe is an excellent baseball player* from the PREMISE that *Joe plays on an excellent baseball team*. Unfortunately, we cannot—to do so would be to form a false link of fallacious REASONING from the PREMISE to the CONCLUSION. For more on this fallacy and others, consult the Wikipedia page on fallacious reasoning.

Additionally, as noted in SPRY, REASONING must be *consistent with current knowledge*. This current knowledge may derive from approved practices within the domain for which the ARGUMENT has been created, it may be based on the received wisdom within the relevant community(ies), or it may employ new ideas from state-of-the-art research that has been deemed to be sufficiently mature to be treated as believable.

We cannot overemphasize the reliance on current knowledge. There is no replacement for it, even though, consequently, it renders the acceptability of knowledge-based REASONING as necessarily variable over time. REASONING that was once considered acceptable may become unacceptable because of changes in knowledge. The aviation domain is full of examples of this phenomenon. An accident happens. The investigation reveals that the causal complex leading to the accident included an event, or an interaction among systems, or an environmental condition, or something else that heretofore had been considered sufficiently unlikely as to be dismissible. From that time forward, whether REASONING is given explicitly or only implicitly, no REASONING will be considered consistent with current knowledge if it permits dismissing the possibility of that event, interaction, condition, or condition occurring.

Relatedly, your PREMISES should all be relevant: they should all, in some way, help to get you to the CONCLUSION of your ARGUMENT. Arbitrary PREMISES, though not inherently damaging¹⁶, make assessing an ARGUMENT difficult, and should be avoided.

5.2.3 Organize Dependence Relations

This section is something of a preparation for the more substantial work the following section requires of the assessor. It may feel less productive, because you're not looking to get a check-mark in this section: rather, an assessor here explores the shape of the ARGUMENT so that when they go to assess it, they are better equipped to do so. It is akin to looking at the map before the start of a long journey—you can always just look as you go along, but it's better to have a sense of where you're going, and how many mountains you will have to cross.

It is possible to proceed to the following step without completing this one, for the ideas that arise in this step will come up again as you complete the “weigh your CONCLUSION” section. You will have to continue consulting your map. That said, this step *can* be done separately, and it can make the following one much easier, so

¹⁶The interested reader is encouraged to ponder whether there may exist specific individual arbitrary PREMISES or combinations thereof that could cause damage beyond just making assessment harder.

we recommend doing it now.

In this step, the focus is on the relationships between PREMISES—what we highlighted in Section 2.4 as the *context* of a PREMISE. Sometimes, a PREMISE can only support the CONCLUSION with help from another PREMISE. An example of this is a conjunction: to support the CONCLUSION “Ms. Scarlett killed Mr. Boddy”, the PREMISE “Ms. Scarlett was in the dining room all evening” is only useful given another PREMISE, “Mr. Boddy was killed in the dining room”, and vice versa. Syllogisms, too, introduce PREMISES that depend on each other: saying that “All men are mortal” only helps us to conclude that “Mr. Boddy is mortal” if we also have the PREMISE that “Mr. Boddy is a man”.

Dependencies occur across most different argumentative forms, and keeping track of them is particularly important considering different amounts of confidence in different PREMISES. Imagine I am yet again trying to argue that “Ms. Scarlet killed Mr. Boddy”. I have high confidence in the PREMISE that “Mr. Boddy was killed in the dining room”, but quite low confidence in the PREMISE that “Ms. Scarlet was in the dining room all evening”. My low confidence in Ms. Scarlet’s whereabouts not only weakens the usefulness of *that* PREMISE to my overall ARGUMENT, but also damages the relevance of the former, such that even though I am highly confident that Mr. Boddy *was* killed in the dining room, that piece of location information has ceased to matter very much.¹⁷

Therefore, the goal of the “Organize” section is to identify and understand the dependencies between PREMISES. When you begin to analyze the PREMISES’ confidence levels, this deeper understanding of their relationships will allow you to assess more effectively how PREMISES’ interactions complicate confidence levels. There is no definitive definition of “done” for this step, because even when you move on from this section, you must continue to think about these interactions. Looking at your map before starting your journey is helpful, but it would be foolish to think that you’ve studied it enough that you can just leave it at the trailhead.

5.2.4 Weigh Your CONCLUSION

By now, an assessor should have completed the preparatory work of confidence assessment by ensuring that the ARGUMENT under consideration is not syntactically or argumentatively flawed, and understanding its shape. In this section, we will loosely outline a method for determining your overall confidence in an ATOMIC ARGUMENT’S CONCLUSION. Though assessing your confidence in the CONCLUSION is, of course, what you are doing throughout SLOW, this is the section where most of the substantive work will occur.

We want to emphasize, here, that the end goal of SLOW is determining what confidence level is engendered by the ARGUMENT—*not* assessing whether it meets

¹⁷Though here, for clarity’s sake, we have used examples of dependencies whereby the PREMISES are *completely* dependent on each other, and failure of one PREMISE dooms the other to irrelevance, dependence is not always this dramatic. If my PREMISES for concluding “Ms. Scarlet killed Mr. Boddy” include “Ms. Scarlet was in the house at the time” and “No one else was in the house at the time”, the former PREMISE is still relevant (though much, much less relevant) even without the latter.

some preset standard. Upon completion, you should have a phrase such as “very confident,” “somewhat confident,” or other such qualitative metrics as your field may demand. The only exception to this is when the CONCLUSION of the ATOMIC ARGUMENT under consideration is the final CONCLUSION: either because the ATOMIC ARGUMENT is not a part of a COMPOUND ARGUMENT, or because it is the all-encompassing ARGUMENT of a COMPOUND ARGUMENT. Then, you will have a standard of confidence you are trying to reach, and you will compare the engendered level of confidence to the standard required.

5.2.5 How to Weigh your CONCLUSION

What you should have in front of you now is a syntactically valid ARGUMENT with valid REASONING which you understand. You should know the confidence levels of each of your PREMISES. With this before you, we recommend the following:

1. *Look at your most confident PREMISES.*

Start by examining what sort of confidence in your CONCLUSION is engendered simply by those PREMISES in which you are most confident. If these were your only PREMISES, how confident would you be in your CONCLUSION?

Which PREMISES are your most confident PREMISES will depend, of course, on the ARGUMENT itself. If you have seven PREMISES in which you are “extremely confident,” and three in which you are “very confident,” then lucky you—start with the seven. If you have two in which you are “very confident” and six in which you are “somewhat confident,” then you would start with the “very confident.” Alternatively, maybe you have one “extremely confident” PREMISE, two “very confident” PREMISES, and four only “fairly confident” PREMISES—in that case, it could make sense to begin with both the “extremely confident” PREMISE *and* the two “very confident” PREMISES. Since this process is, as we said, meant only as a tool, there is no absolute guideline here. It’s all relative to your ARGUMENT.

Recall from the “Organize Dependence Relations” section that some PREMISES are only relevant in combination with other PREMISES. In this step, your “very confident” PREMISES might not yet be accompanied by the counterparts that make them relevant, so they may lend very little confidence (if any) to the CONCLUSION right now.

This might seem counterintuitive—shouldn’t we consider PREMISES that belong together, together? Indeed, we ultimately will do so. But beginning here is useful because, when two PREMISES depend

on each other like this, it is all too easy to be blinded by high confidence in one of them, and miss how low confidence in the other seriously undermines its helpfulness to the CONCLUSION. By initially separating them, we are able to see the ultimate interplay of the confidence levels more clearly.

By the end of this step, you should know the confidence level that you would have assuming that your *only* PREMISES were the ones in which you are most confident: i.e., how far would that truncated ARGUMENT get you?

2. *Add in some more PREMISES.*

Now, you will consider PREMISES in which you are less confident, and to what extent these PREMISES—such as they are, i.e., in their not-so-confident-form—still increase your confidence in the CONCLUSION.

As before, there is no clear answer as to what the confidence level of these ‘new’ PREMISES should be. It is dependent on your ARGUMENT, and even then, there might be several ways to go about choosing which PREMISES are next.

You may find it fruitful to next draw into your ARGUMENT your next confident PREMISES—so if you have only been assessing “extremely confident” PREMISES, you might next look at “very confident” PREMISES. Alternatively, it might make more sense, within the context of your ARGUMENT, to now look at the PREMISES which are in dependence relations with your extremely confident PREMISES. Neither of these answers is wrong.

It may seem that once you get into the less confident PREMISES, they will necessarily be less helpful in engendering confidence in your CONCLUSION—you might even ask yourself what the point is of examining these PREMISES at all. It is true that any PREMISE is more helpful when it is more confident. However, it does not follow that the more confident PREMISES are always the most helpful. If, in trying to prove that Ms. Scarlet killed Mr. Boddy, I am *extremely* confident that Ms. Scarlet had a motive, but only *very* confident that Ms. Scarlet was caught on fifteen security cameras in the very act of the crime, it does not seem unreasonable that the latter would increase my confidence in the CONCLUSION more than the former.¹⁸

¹⁸This idea—that some PREMISES in themselves have better potential to guarantee the CONCLUSION—is what we referred to as *content* in section (3).

With this in mind: keep fighting the good fight and considering PREMISES, even the PREMISES that are not so confident.

3. *Continue adding.*

Keep adding in PREMISES to your consideration and examining the extent to which they engender more confidence in your CONCLUSION until there are *no more* PREMISES. Again—the order in which you add PREMISES to your consideration may happen in different ways.

What is important to remember is that you cannot stop until you have considered all PREMISES. The goal of this process is to discover the level of confidence that the ARGUMENT in its totality engenders in your CONCLUSION. To stop before you have considered all PREMISES would run the risk of underestimating your confidence level. Stopping before you have considered all PREMISES, then, is only appropriate in the case that this CONCLUSION is the final one, and you have reached your required level of confidence. However, *even then*, it seems better to consider all PREMISES, just in case you have missed or misunderstood a dependence relation.

That said, given correct application of SLOW, your confidence in the CONCLUSION should never decrease as you consider more PREMISES. Since all your PREMISES have been deemed as helpful for reaching your CONCLUSION, a PREMISE *decreasing* your confidence in the CONCLUSION should be cause for concern. Your confidence in the CONCLUSION may, however, stay the same as you consider a new PREMISE, rather than increasing. This could happen for several reasons: either (i) the PREMISE was never truly relevant, (ii) the PREMISE itself has such little confidence that it can't engender confidence in the CONCLUSION, or (iii) the PREMISE's relevance has been undercut because its necessary counterpart lacks sufficient confidence to support it.

It might also happen that a PREMISE does not *yet* increase your confidence in the CONCLUSION—when a PREMISE is in a dependence relation and you have not yet looked at its counterpart, its helpfulness will remain in limbo, for you cannot say how much confidence it engenders until it is reunited with its friend(s).

That said, at the end of this step, you should have determined your confidence level in the CONCLUSION.

5.2.6 Further Notes on Weighing Your Conclusion

In this section, we have provided a loose method for making your ultimate confidence assessment. But this is *not the only method* of completing the “weigh” portion of SLOW. The goal of this final step is discovering how much confidence the ARGUMENT engenders in the CONCLUSION, and this could be done any number of ways. We have provided a structure because it is useful to break down the assessment process, and there seem to be certain elements of assessing confidence, such as dependence relations, which can introduce complexity. Breaking up the ARGUMENT into bite-size pieces thus seems fruitful. Our structure is centered around highlighting the confidence levels of PREMISES; however, so long as you keep sight of your confidence level in each PREMISE, and consider all PREMISES, the way in which you realize this goal is flexible.

For instance, your ARGUMENT might be brief enough such that you can just read through it and assess it all in one bite. If this is possible, then it is not absolutely necessary to do the above-detailed process by any means. Alternatively, you might prefer to focus on dependence relations, examining how much confidence each small cluster of interrelated PREMISES engenders in the CONCLUSION. Or, perhaps, you may begin with those PREMISES which seem like they have the most potential to guarantee your CONCLUSION—the *Ms. Scarlet on fifteen security cameras* type—and see whether they have high confidence. All this is to say: there is no one objectively right way to go about this final step.

That said, we do like our structure in particular for three reasons. First of all, since you already have a clearly specified confidence level for each PREMISE, it seems like the easiest way to organize PREMISES, rather than getting possibly muddled in which ones seem potentially helpful, or trying to disentangle complicated dependence relations. Secondly, as we said before, it seems easy to be blinded by high confidence in one side of a dependence relation such that you lose sight of lack of confidence in another side, and our method helps to prevent that. Lastly, our method seems markedly efficient in comparison with others we can think of. Since the PREMISES in which one is highly confident will *typically* (though not necessarily) engender a large amount of your confidence in the CONCLUSION, it follows that you will more quickly get an accurate sense of your ultimate confidence.

We also should note that there is a case in which our process does *not* seem particularly useful. When you have an ARGUMENT in which you are equally confident in all of the PREMISES—whether this is because all of your PREMISES are terminal PREMISES, or simply because they all happen to have the same level—our method does not guide you in breaking down the ARGUMENT into bite-sized pieces. It still seems fruitful to break an ARGUMENT down into bite-sized pieces: humans can only consider so many PREMISES at a time. In this case, however, the way in which you break down the PREMISES does not seem as important, since the standards of confidence are very simple to keep track of. For ARGUMENTS of this type, we leave the order of consideration to the assessor to determine.

5.3 STAID

In the previous section, we explained a method for evaluating **ATOMIC ARGUMENTS**, called **SLOW**. In this section, we will provide a method for evaluating **COMPOUND ARGUMENTS** which integrates **SLOW**.

The guiding idea behind this method, called **STAID**, is that you cannot accurately evaluate a non-deductive **ARGUMENT** when you do not know your confidence level in the **PREMISES**. Thus, the method is meant to ensure that you do not lose information about confidence levels as you evaluate the **ARGUMENT** in its entirety. **STAID**'s strategy for retaining this information is to start at the 'bottom'—the terminal **PREMISES**—and work its way 'up'; that way, you only ever assess **ARGUMENTS** whose **PREMISES** have known confidence levels, and no information is obscured.

Our procedure is made up of five basic steps:

1. Assess Overall **Structure**
2. Identify & Assess **Terminal PREMISES**
3. Identify & Assess an Unlocked **ATOMIC ARGUMENT**
4. **Iterate**
5. **Declare COGENT or non-COGENT**

5.3.1 Assess Overall Structure

This is a preliminary step in assessing a **COMPOUND ARGUMENT**, which you undertake before getting into confidence levels, **ATOMIC ARGUMENTS**, or anything else pertaining to the actual meaning of the **ARGUMENT**. This step is purely about inspecting the structure for a particular aspect that could endanger the **COGENCY** of this **ARGUMENT** right off the bat: something we are calling a *cycle*.

A cycle occurs when two or more defended PREMISES are used in one another's ATOMIC ARGUMENTS, such that each confirms the other. For more clarity, consider the following example: you are trying to prove that Ms. Scarlet murdered Mr. Boddy in the dining room with a candlestick. The first thing you do is to split this **ARGUMENT** into **PREMISES**: you want to show that (a) Ms. Scarlet killed Mr. Boddy; (b), Mr. Boddy was murdered *in the dining room*; and (c), Mr. Boddy was murdered *with a candlestick*. Lucky you—or so you think—you stumble upon the following piece of evidence: (d), the dining room is chock-full of candlesticks, and no other room has a single one.

You decide to use this **PREMISE** in support of (b) and (c), since it seems relevant to both. However, for all those candlesticks in the dining room to be relevant to (b), we would have to know (c), that Mr. Boddy was murdered with a candlestick. Similarly, for all the candlesticks in the dining room to be relevant to (c)'s defense, we would have to know (b), that Mr. Boddy was murdered in the dining room.

Consequently, you end up using your two defended **PREMISES** (b) and (c) in one another's defense, aided by the **PREMISE** (d). When you move into assessing

an ARGUMENT that has a cycle like this, the two PREMISES will end up increasingly affirming each other, making it seem like you can be much more confident than you actually are.

Fortunately, STAID is quite literally *impossible* to complete when cycles remain in the ARGUMENT. You will stumble upon them along the way. Therefore, if things *do* get impossible as you go through the process, try and step back, and ask yourself whether you have found a hidden cycle.

Nevertheless, it's no fun to be stymied in the middle of the process, so try to get it all done now: *look for cycles*, and if you find one, declare the ARGUMENT to be non-COGENT and hand it back to the creator for rewriting.

5.3.2 Identify and Assess Terminal PREMISES

Terminal PREMISES, as we have said, are PREMISES which are *not* the CONCLUSIONS of their own local ARGUMENTS. These are your argumentative building blocks, your foundations. Naturally, strong foundations are desirable. In this step, then, you are *identifying* which PREMISES are terminal and then assessing each of them to ensure that they are strong.

Identifying terminal PREMISES is simple: first, determine which of your PREMISES are not supported by a local ARGUMENT. Once you have identified all your terminal PREMISES, you must assess them. To be considered valid, a terminal PREMISE must fall into one of the following categories, which correspond with PREMISE categories (a, b, d, f) in SPRY:

- (a) the terminal PREMISE expresses a proposition that is 'universally' accepted as true
- (b) the terminal PREMISE expresses a proposition that is accepted as true within the relevant domain
- (c) the terminal PREMISE is supported by external artifacts that are accepted within the domain as being sufficient to establish its truth
- (d) the terminal PREMISE is an assumption accepted by all stakeholders and clearly identified as such

For more description of these categories, please refer to the "Primer on ARGUMENT Assessment" (PASS), cited in the bibliography.

5.3.3 Assess an Unlocked ATOMIC ARGUMENT

Now that you have assessed all of your terminal PREMISES, you will have at least one **unlocked ARGUMENT**.¹⁹ Recall that an unlocked ARGUMENT is an ARGUMENT wherein all the PREMISES have known confidence levels, i.e., all the PREMISES are **assessed PREMISES**. Only when you know the confidence levels of your PREMISES can you accurately assess an ARGUMENT.

¹⁹If you have no unlocked ARGUMENTS, then you have indeed stumbled upon a cycle.

You might in fact have multiple unlocked ARGUMENTS. If this is the case, simply select one, and assess it using SLOW. Doing so successfully will result in the CONCLUSION of the ARGUMENT going from an **unassessed defended PREMISE** to an **assessed defended PREMISE**, perhaps unlocking more ARGUMENTS.

5.3.4 Iterate

Pick another **unlocked ARGUMENT** and assess it using SLOW. Then pick another, and do it again. You will continue to unlock ARGUMENTS as you continue to assess, because you will turn more and more **unassessed PREMISES** into **assessed** ones.

If the ARGUMENT has no cycles, this step will run smoothly. You will be able to continue picking unlocked ARGUMENTS and assessing them, thereby unlocking more ARGUMENTS, until you reach the final ARGUMENT: the ATOMIC ARGUMENT that brings you to the ultimate CONCLUSION.

5.3.5 Declare COGENT or non-COGENT

In assessing this last ATOMIC ARGUMENT with SLOW, you will determine the engendered confidence level in your ultimate CONCLUSION. All that is left to do is to compare this level with the standard required for declaring COGENCY. If the confidence engendered by the ARGUMENT reaches or exceeds this standard, then mark it **COGENT**. If it falls short, mark it **non-COGENT** and hand it back to its creator.

6 Further Directions

We have now finished our account of STAID as an alternative method of assessing ARGUMENTS. In this section of the paper, we turn to next steps, with brief commentary on the method and its situation in regard to other strategies to justify safety.

6.1 Possible Problems

There are certainly many objections that one may raise in response to STAID, and as more people read this paper, more objections will come to light. However, two primary concerns come to mind, so we will spend a moment addressing them.

First, readers may be concerned that this process is complicated, indeed, even more complicated than SPRY and iTest. This is a correct assessment; as SLOW and STAID require a deeper understanding of the ARGUMENT's structure as a whole and the different amounts of confidence engendered by different ATOMIC ARGUMENTS, they necessarily require more time and effort from the assessor. Sadly, we do not live in a utopia of infinite time and energy to assess ARGUMENTS, and thus, efficiency is truly important. In situations where leaking confidence is not a concern (for example, if you have a profusion of evidence in which you are absolutely confident or nearly so, or your ARGUMENT is very brief), it is a perfectly reasonable decision to use iTest.

That being said, just because SLOW and STAIID are *more* complicated does not mean that they are *too* complicated. In some ways, it is very simple: you are doing nothing more than breaking down the ARGUMENT and comprehending its structure. Particularly if the ARGUMENT is well-written and not convoluted, then applying this system should not be too hard, and seems a reasonable thing to require of an assessor.

The second problem readers might identify is the possibility that assessors could import their own biases when assessing ARGUMENTS using SLOW and STAIID. Our procedure depends heavily on an assessor's judgment, particularly in the "weighing" portion of SLOW, and thus might seem particularly open to this issue. This, too, is a fair concern.

We might first note that this concern holds for *any* system of assessing ARGUMENTS, and, in fact, for any system for justifying confidence. Biases are always present, and the best way to face them is to recognize where they might arise. As such, being aware of the possibility of importing biases is in itself a defense. Assessors should be able to circumvent this problem to the degree that it can be circumvented in any mode of justifying confidence so long as they go in with good faith, and assess ARGUMENTS SLOWly and STAIIDily.

6.2 A Dead End

One impulse when working with ARGUMENT assessment is quantification. This is a sensible impulse, particularly if you are concerned about importing your own beliefs: to some²⁰ people numbers *seem* more trustworthy and objective than simple judgments.

One reason quantification may be very tempting as there seem to be many patterns in how a lack of confidence aggregates across ARGUMENTS. There are many fields working on this phenomenon under names like "Bayesian Belief Networks" and "Probabilistic Logic". Though they have found (and proven!) many interesting theorems, and indeed some concerning how uncertainty in PREMISES is transferred to CONCLUSIONS, these mathematical tools are simply not strong enough to predict confidence. They rely on assigning precise probabilities to PREMISES (which, in most real-world cases, is implausible or impossible to do with any accuracy), and even then can only give you a range of possible probabilities that your CONCLUSION is true.²¹

²⁰Perhaps many, but definitely not all. One of the co-authors of this paper is quite fond of needling number-loving colleagues: "If have you to put a number on it, you don't understand it."

²¹In these remarks, we are largely thinking about Adam and Levine's "On the Uncertainties Transmitted from Premises to Conclusions in Deductive Inferences," 1975 [7]. This paper proves that there is an upper and lower bound on the "uncertainty" of a CONCLUSION based on the "uncertainty" of the PREMISES in a deductive ARGUMENT. This is promising and relevant work, but there are multiple elements that make it critically unprepared to deal with something like confidence in OPRAs. It treats uncertainty as "probability that the statement is false", only deals with deductive ARGUMENTS, and *still* is not able to precisely calculate the uncertainty of the CONCLUSION. Perhaps one day these theorems will develop to the requisite particularity and precision to help in the assessment of OPRAs, but today is not that day. Readers interested in learning more about the state of current confidence quantification tools may also want to take a look at [8,9].

Such methods also struggle to formalize the idea that different PREMISES relate differently to the CONCLUSION and to each other; by stripping PREMISES of their context, they obscure these important interrelations which we have highlighted above.

Instead of trying to strip down confidence in the COGENCY of ARGUMENTS into numbers, we should work to retain as much substance as possible. Competent people are able to think through ARGUMENTS quite well, and mathematical theorems can not be made precise enough to judge justified confidence in the sorts of CONCLUSIONS relevant to us, a particularly insidious problem because numbers have the appearance of absolute precision.

6.3 An Open Path

One path that seems particularly fruitful to us is reconciling SLOW and STAID with SPRY and iTest. Here we provide a few preliminary thoughts as to how this might go.

SLOW, being based on SPRY, could potentially be built into SPRY rather than remaining a separate test. The first step is essentially the same. The second and third steps are similar enough that the steps of SLOW could be reworked to closer reflect the order of SPRY, and, in some places, directly coincide. However, there is one intractable difference between SPRY and SLOW: the former results in a *yes* or *no*, while the latter results in a confidence level. Thus, embedding SLOW in SPRY would require significant changes to the final step of SPRY.

The difference in end result also motivates the “Get Organized” step in SLOW: because it is more complicated to come up with a particular confidence level, it is important to *slow down*, and have support in place for the final step. If embedding SLOW in SPRY, then, it would also be important to retain this support system in a meaningful way.

At first glance, STAID and iTest seem far more incompatible; their descriptions do not align as obviously as those of SPRY and SLOW. However, iTest has a lot of flexibility in how it is performed, and this flexibility renders STAID and iTest quite compatible. Since one can choose any order of assessing local ARGUMENTS one wishes, STAID could be one manner of performing iTest: a manner distinguished by limiting one’s order of assessment to the demands of “unlocked” and “locked” ARGUMENTS. When performing STAID, you do not have to break any of the rules of iTest, you simply have to abide by additional constraints. What is crucial is that their central concepts are precisely the same: iterate through many atomic ARGUMENTS until the COGENCY of the final ARGUMENT is determined.

These are just preliminary thoughts on the possibility of such reconciliation. To actually make these processes compatible or embed one in the other, a fuller exploration and a new document would be required. We decided in this document to make a separate system in order to emphasize the differences we are introducing and their importance, but eventually reconciling these systems seems feasible and practically important.

7 Conclusions

The purpose of this paper has been essentially twofold: first, to present a concern about assessing ARGUMENTS based on confidence, and second, to provide an assessment procedure which takes this concern into account.

This concern is that, throughout a COMPOUND ARGUMENT, lack of confidence can aggregate in ways that are difficult to predict. Non-deductive ARGUMENTS are tricky because even when the REASONING is valid, it leaves room for doubt. The lack of confidence inherent in the REASONING coalesces with the lack of confidence inherent in PREMISES. And, over the course of a COMPOUND ARGUMENT, this happens many times over, thus accumulating. Therefore, assessment for each ATOMIC ARGUMENT must consider the confidence levels of each of its PREMISES and the particularities of its REASONING.

With SPRY and iTest, this is made difficult because ATOMIC ARGUMENTS can be assessed in any order. It is thus not possible to include the confidence levels of the PREMISES in one's assessment of the ATOMIC ARGUMENT. The iterative process which allows PREMISES to be used provisionally until they are deemed "acceptable" or "not acceptable" does not account for the fact that PREMISES might fall in between and still be relevant for the CONCLUSION.

In the above, we have provided a different procedure, SLOW and STAID, which not only requires assessment of each ATOMIC ARGUMENT but retains information about confidence throughout. This requires the assessor to bring this information to bear at each level, continually reassessing what sort of confidence they *have*, instead of what sort they *need*. We anticipate that there will be further questions or problems with this procedure, but we BELIEVE this method of ARGUMENT assessment is a worthy endeavor.

References

1. Holloway, C. M.; and Wasson, K. S.: A Primer on Argument. White paper, National Aeronautics and Space Administration, Hampton, VA, USA, June 2021. URL <https://ntrs.nasa.gov/citations/20210019993>.
2. Holloway, C. M.: The Friendly Argument Notation (FAN). Technical Memorandum NASA/TM-2020-5002931, National Aeronautics and Space Administration, Hampton, VA, USA, June 2020. URL <https://ntrs.nasa.gov/citations/20205002931>.
3. Holloway, C. M.; and Wasson, K. S.: A Primer on Argument Assessment. White paper, National Aeronautics and Space Administration, Hampton, VA, USA, October 2021. URL <https://ntrs.nasa.gov/citations/20210022807>.
4. Holloway, C. M.: Understanding the Overarching Properties. Technical Memorandum NASA/TM-2019-219650, National Aeronautics and Space Administration, Hampton, VA, USA, July 2019. URL <https://hdl.handle.net/2060/20190029284>.

5. Wasson, K. S.; and Holloway, M.: An Introduction to Constructing and Assessing Overarching Properties Related Arguments (OPRAs). White paper, NASA Langley Research Center, January 2022. URL <https://ntrs.nasa.gov/citations/20210025425>.
6. Rock Hunters; McCardel, F.; Holloway, C. M.; Wasson, K.; McDonnell, N.; Graydon, M.; Peña, A.; and Lehman, S.: Towards a Coherent View of Evidence in Safety Assurance. Technical Memorandum NASA/TM-20230003336, National Aeronautics and Space Administration, Hampton, Virginia, April 2023. URL <https://ntrs.nasa.gov/citations/20230003336>.
7. Adams, E. W.; and Levine, H. P.: On the uncertainties transmitted from premises to conclusions in deductive inferences. *Synthese*, vol. 30, no. 3, 1975, pp. 429–460. URL <https://doi.org/10.1007/BF00485053>.
8. Graydon, P. J.; and Holloway, C. M.: An Investigation of Proposed Techniques for Quantifying Confidence in Assurance Arguments. Technical Memorandum NASA/TM-2016-219195, National Aeronautics and Space Administration, Hampton, VA, USA, May 2016. URL <http://hdl.handle.net/2060/20160006526>.
9. Graydon, P. J.; and Holloway, C. M.: An Investigation of Proposed Techniques for Quantifying Confidence in Assurance Arguments. *Safety Science*, vol. 92, Feb. 2017, pp. 53–65.