A Confidence System for Solving Real-World Problems with Argumentation

Laura Rassbach de Vesine

University of Colorado at Boulder

Technical Report CU-CS 1060-10

February 2010

1 INTRODUCTION

Reasoning with pure logic, where once a conclusion has been reached it is unassailable, is a luxury available to AI researchers only rarely. Instead, for real problems many researchers find that some notion of doubt and defeat must be represented to have any hope of modelling real human reasoning. Argumentation has shown significant promise at addressing these problems. However, few argumentation systems have been implemented to solve real problems. As a result, the confidence and defeat systems of many existing argumentation architectures are overly simplistic and do not address all of the issues that arise in solving practical problems. Choosing a good representation and combination system for confidence is crucial to solving real problems. Better representations of confidence should allow us to more accurately and completely model human reasoning with less effort. On the other hand, simpler systems allow us to focus on the problem of interest, rather than continually adding more layers of complexity to the confidence system.

I have found that representing confidence as a two-dimensional vector strikes this balance well. While this representation is not as simple as a single unit of confidence, it is significantly more expressive for representing confidence in real problems. However, this two-dimensional representation makes comparing confidence in opposing conclusions more difficult.

This paper presents an argumentation system that addresses these problems in detail. I have implemented this system using Calvin, an argumentation framework based on Krause et al.'s LA [Krause et al., 1995], and used it to address two real-world problems. Calvin is successful at solving a reasoning problem in cosmogenic isotope dating, a branch of geology, and choosing whether to open bridge hands. These two problems are quite different, but Calvin's confidence system captures the reasoning of experts in both problems, indicating the broad applicability of this confidence framework to human reasoning problems.

2 A CONFIDENCE SYSTEM FOR REAL PROBLEMS

2.1 Motivation

It is evident even to the casual observer that some arguments carry greater weight than others. However, precise comparisons between distinct arguments are not always easy to accurately perform. For example, some arguments about whether the milk in the refrigerator has spoiled might be:

- (1) This milk is one week past its expiration date. Old milk spoils; therefore, the milk has gone bad.
- (2) I have not purchased milk recently. Old milk spoils; therefore, the milk has gone bad.
- (3) This milk is still a normal color. Spoiled milk eventually changes color. Therefore, the milk has not gone bad.

Clearly (1) and (2) are quite similar arguments, sharing the same root rule. In fact, many systems would derive these arguments as a single tree with two branches. However (1) is a stronger argument for the milk having gone bad because it draws on empirical observations of the actual milk rather than general information about when milk was last purchased. This issue is sometimes handled in argumentation systems by referring to the *specificity* of arguments, with more-specific arguments carrying more weight [Elvang-Gøransson et al. 1993]. However, (3) seems to contradict this choice of weighting for Calvin: although it refers to a specific observation of the milk, it is a weaker argument than (1). Furthermore, the relationship between (2) and (3) is surprisingly difficult to quantify. Finding a computational way to describe the relative strengths of these three arguments, one that preserves the intuitive relationships between them and the fact that they are somehow intuitively difficult to compare, is a surprisingly difficult problem that is easy to overlook without implementing a system to solve a real problem.

2.2 Related Work

I have drawn on several other logical argumentation systems for my definition and use of confidence. These include [Prakken 2005], [Cayrol and Lagasquie-Schiex 2003], [Morge and Mancarella 2007],

and [Farley 1997]. [Prakken 2005] defines a system of accrual for inference-based logics (in which he includes LA) that, in a bottom-up manner, first combines the evidence for and against each particular conclusion, then weighs competing evidence, and finally proceeds to use the resulting 'winner' in further arguments. Furthermore, this formalism includes the notion that an argument is only as strong as its weakest link and that multiple weaker arguments combine to create a single stronger argument. This procedure for the combination of confidence seems promising since it is powerful and yet computationally simple, but Prakken defines neither a format for confidence nor a method of comparison.

[Cayrol and Lagasquie-Schiex 2003] define a method for "gradual" ranking of arguments based on the number of defeaters, the number of defeaters of defeaters, etc. As in other logical/inference-based argumentation systems, the complete set of possible arguments is formed and then evaluated. Rather than requiring the explicitly-defined, absolute defeat of a *rule*, a specific *conclusion* is gradually defeated by having more attacking arguments than defending ones. However, [Cayrol and Lagasquie-Schiex 2003] does not consider the absolute weights of individual arguments. Thus, all arguments have the same weight, restricting the richness of the system and making it essentially impossible to adequately solve the spoiled milk question from above.

[Farley 1997] uses argumentation as a method for performing qualitative simulation in the presence of conflicting indications. His system allows three different modes: accept all arguments, accept the side with more arguments, and accept all defeaters. In the first mode, no conclusion can be defeated. Every conclusion is accepted if there is any undefeated argument for it. This mode is intended to encourage experimentation. In the second mode, all arguments have the same strength, and the side with more undefeated arguments 'wins.' Finally, in the third mode, a conclusion is accepted if and only if there is an undefeated argument for it, and all arguments against it are defeated. This mode is intended for skeptical reasoning. Three classes of arguments are defined, with a hierarchical defeat mechanism between them. Arguments are explicitly defeated only when they are attacked by an argument with a higher type in the hierarchy; cases with arguments of the same strength are handled by the system modes.

[Cayrol and Lagasquie-Schiex 2003] and [Farley 1997] have effectively opposite weaknesses from the perspective of practical confidence. While [Cayrol and Lagasquie-Schiex 2003] allows for the partial defeat of arguments, there is no measure of individual argument strength. On the other hand, [Farley 1997] contains arguments with different weights but does not allow for partial defeat. In the example above with spoiled milk, (3) partially defeats the conclusion from (1) and (2) that the milk has spoiled, but does not do so entirely. While it is insufficient evidence to defeat the conclusion alone, it can contribute to the overall evidence required to convince us that the milk is still good in the face of its age.

[Amgoud et al. 2005] discuss a decision support system that uses a multi-vectored approach to comparing arguments. Their system weighs arguments in favor of a specific decision on the basis of three dimensions: certainty of the knowledge used to form the argument, degree of satisfaction of the required criteria, and importance of goals. They use a "psychologically valid" method of weighing arguments against each other that, in the presence of "extreme" criteria, chooses the decision with the weakest negative argument. In contrast, when there are no extreme criteria, the system selects the conclusion with the strongest positive argument. Although this division of confidence seems promising for the problem of practical confidence in logical propositions, the conceptual difference between the strengths of goals and the strength of evidence prevent this decision principle from being useful in actual practice. [Morge and Mancarella 2007] also discusses the importance of a multiple element confidence vector including the certainty of knowledge and the importance of various goals to selecting a decision via argumentation, but they assume that these criteria are extra-logical and provide no mechanism for comparing confidences. Making the criteria for confidence comparison extra-logical forces the user to cope with significant complexity that might be better handled by the argumentation system. confidence are *not* extra-logical, and it should therefore be able to compare and combine them.

2.3 Representing Confidence

The crucial insight for developing a confidence system that is both simple enough to be intuitive and rich enough to adequately express problems like the question of spoiled milk was the realization that not only can specific *evidence* be trivial or critical, but the *knowledge* used to connect the evidence to the conclusion is also of variable quality. This choice to define confidence with two dimensions instead of one makes it clear why one argument is better than another, something experts handling real problems know when they make their arguments. In the arguments about spoiled milk given above, (1) uses both higher-quality evidence and higher-quality knowledge than the other two arguments. (2) uses high-quality knowledge but only moderate-quality evidence. (3) uses high-quality evidence but much lower-quality knowledge.

This insight led us to use a system of confidence based on two-element vectors containing a restricted set of values. The first element of the vector, 'applicability,' represents how closely the actual evidence matches the prototypical situation the knowledge is drawn from. This system uses three levels of applicability: partly, mostly, and highly. The choice of three levels of applicability is drawn from possibilistic logic [Farreny and Prade 1986], which often uses real numbers but effectively reasons in terms of 0, 1, and 'between 0 and 1,' and is thus representable with three qualitative values. For each level of applicability, an argument either supports or refutes the conclusion at hand, effectively implementing a 'not' operator. As an example of how applicability works in practice, consider a case where we know that a positive correlation between *x* and *y* is sound evidence of conclusion *z*. In order to apply this knowledge, first calculate the correlation coefficient between *x* and *y*. The correlation may be quite strong, moderate, or very weak: thresholds for these categories are contained in the rule and are chosen as part of the knowledge acquisition process. If the coefficient is quite high (very near 1) then the knowledge is highly applicable in this case. If it is low but still positive (say 0.2), then the knowledge is only partly applicable: the positive correlation may simply be a statistical artifact, and not real evidence. If the correlation coefficient is slightly negative, it is partly applicable evidence against *z*, by the same reasoning. A correlation coefficient very near -1 would be highly applicable evidence against *z*.

This system of confidence also assigns applicabilities in cases where the knowledge involves a binary quantitative comparison. Because quantitative observations are generally noisy, it makes sense to have more confidence in an observation when the values in the relation are farther apart: when a noisy observation is quite close to some cutoff value, it is more likely that the actual relation to the cutoff is reversed (i.e., that the noise has moved the observation across the cutoff). More concretely, for an rule like "*x* < 50 => *y*," the applicability is higher with a value for *x* of 20 than a value of 49. Quantifiers such as 'for-all' and 'there-exists' can be easily handled by selecting the most-true (for 'there-exists') or least-true (for 'for-all') applicability value among the quantified entities. That is, if the rule asks whether *any one of x*, *y*, or *z* is < 50, the input with the lowest value is compared to the threshold. If the question is whether *every one of x*, *y*, and *z* is below that threshold, the input with the highest value speaks for the whole group.

The second element of confidence, 'validity,' represents the quality of knowledge involved in the conclusion. A specific validity value is assigned to every rule as a measurement of the strength or trustworthiness of the knowledge expressed in that rule, from a gut feeling to a universally accepted theory. Validity has four possible values: plausible, probable, sound, and accepted. Other argumentation systems use about this many general classes of argument, including [Elvang-Gøransson et al. 1993]. These four terms cover the range of qualities of knowledge used by experts so far in my experience. Plausible knowledge denotes gut reasoning with little real support, e.g. a preference for conclusions that use more of the input data in their reasoning. Probable knowledge is based on repeated experience but without significant experimental support. Sound knowledge is used by virtually all experts in a field and is well supported by existing data, usually including experimental data. Examples might be a rule that says that when a plant is green it is healthy, or a rule that milk is good when it has a normal smell. While these rules do not always hold true, they are generally true and contain significant argumentative force. Accepted knowledge would be a widely accepted theory or mathematical near-certainty. The theory of gravity is an example of accepted knowledge.

2.3.1 Using Confidence

To judge the relative and absolute strengths of arguments using this system of confidence assignment, we need to manipulate confidence values in two distinct ways. The first operates along a single chain of reasoning: milk is more likely to have gone bad when it is old; this milk was purchased some time ago and is probably old. Intuitively, it makes sense to choose the validity of the least-valid rule for the overall conclusion, so that the chain is only as strong as its weakest link. Applicability is 'created' by the direct use of observed evidence. In this case, how long ago the milk was purchased, compared to how old milk usually gets before it goes bad, determines the applicability. This occurs at the leaves of an argument tree. The lowest—that is, least-true or most-false—applicability is used for the conclusion of any rule that uses an 'and' conjunction, and the highest for rules using 'or.' Rules may also lower or raise the applicability of knowledge passed through them when they are applied. This is to handle situations where an observation is not specific to the knowledge being applied, as in argument (2) at the beginning of this section.

The second and more-complicated use of confidence occurs when a number of different chains of reasoning are all applied to the same conclusion (because an argument is a *collection* of trees), such as the conflicting arguments about spoiled milk at the beginning of this section. This can be a complicated problem to solve because a chain of reasoning supporting the conclusion might have higher validity but lower applicability than a chain of reasoning refuting the conclusion. Furthermore, there are often several independent chains of reasoning both supporting and refuting the conclusion, each with its own confidence level. Using the method discussed in [Prakken 2005] we can assign confidence in two stages, first locally up a single chain of reasoning and then globally across many chains of reasoning arguing for the same conclusion.

In determining how to weigh different confidence levels against each other, I followed several general guiding principles. First, the level of validity should be more important than the level of applicability: reasoning about plausible scenarios should not have more weight than the use of sound theories. As a practical example, no matter how green the moon looks from earth, we do not believe that it is made from green cheese in defiance of much closer observations by others. However, in the same way that sufficient circumstantial evidence can carry the same weight as one piece of direct evidence, enough arguments at a lower validity level should eventually result in an argument at a higher validity. This follows Amgoud *et al*.'s [Amgoud et al. 2008] principle that if the quantity of support for an argument increases, the quality of the support increases. Returning to the example of the moon, imagine that not only is it quite green, it smells faintly of green cheese and is noticeably different in shape than it was last night. While this may still be insufficient evidence to override the long-standing scientific consensus that the moon is made of rock, these three elements together make a more persuasive case than any of the individual components. It is important for applicability to also affect the resulting confidence because it represents how closely the current problem matches the knowledge being used for reasoning. Finally, in cases where the evidence is 'tied' for accepting or rejecting a conclusion, weakly reject the conclusion, based on the notion that in science one is more likely to disprove a theory than to prove it.

To determine the overall confidence in a conclusion from a collection of argument trees, first aggregate lower-validity confidences in groups of three into higher-validity confidences. This size of group simply seems to work well in practice; it is not based on some deep insight (results are discussed in detail in Section 4). Then, if the highest-validity confidences for and against the conclusion are at least two levels apart, the highest-validity confidence is returned intact as the overall confidence: it is judged sufficiently strong to completely override the weaker rebutting evidence. A difference of two levels of validity implies a huge difference in overall confidence strength—it is the difference between a logical tautology and a statement such as 'a warm winter makes my garden more likely to be eaten by bugs.' In contrast, a single level of difference in validity is less drastic, for example the difference between the preceding statement about bugs and a statement that 'if my plants do poorly it is plausible they have been attacked by insects.' The resulting confidence in other situations is illustrated in Figure 1 and Table 1. Figure 1 indicates which confidence 'wins,' that is, is the overall confidence in the conclusion. However, when the competing confidence is close to the 'winner,' the weighting system reduces the

overall confidence in the conclusion according to how close the two competing confidences are. Table 1 shows the possible ranks of confidence reduction and what situations they apply to. To combine more than two opposing confidences, simply apply the figure and table iteratively.

To see how this method of confidence combination preserves nuances while arriving at reasonable conclusions, consider a simple example. You are concerned that a gallon of milk has spoiled and you are using the system given in Figure 1 and Table 1. Only two observations are available to you: how the milk smells and how many days, if any, it is past its expiration date. Clearly in such a situation you will have a fair idea of how accurate your sense of smell is and how accurate milk expiration dates usually are, allowing you to judge the overall validity of the conclusions you reach from your observations.

Normal smell, 1 day expired: In this case, the milk smells entirely normal, which is highly applicable evidence that unifies with the knowledge that normal-smelling milk is still good. Your sense of smell about milk is generally a quite accurate indicator of whether it is good, so you can draw the conclusion that the milk is good with high applicability and sound validity. On the other hand, milk's expiration date is also usually an accurate indicator of whether it has gone bad: in fact, you feel that it is about as valid an indicator as your own sense of smell. However, since the expiration date is recent, the conclusion that the milk has gone bad is only partly applicable with sound validity. Referencing Figure 1, we find that when the validities of the competing 'for' and 'against' confidences are equal and the applicability of the 'for' confidence is higher than that of the 'against' confidence, as in this case, we select the 'for' confidence. Then, referencing Table 1 we find that since the validities are equal and the applicability of the winning confidence is two levels higher, we reduce the applicability of the winning confidence by one. Overall, then, we have a mostly applicable soundly valid argument that the milk is still good. This confidence level seems to accurately express the slightly-cautious conclusion that the milk is in fact still drinkable.

Mostly normal smell, 2 weeks expired: In this case, you are less confident that the smell is normal, yielding only mostly applicable sound validity evidence that the milk is still good, and the expiration date is far past (highly applicable). From Figure 1, since the validities of the two confidences are equal, we take the one with the higher applicability: in this case, the argument against the milk being good. Referencing Table 1 again, we find that for equal validity and an 'against' applicability one level higher, we must reduce the overall final confidence by 2 levels of applicability, yielding a partly applicable sound argument that the milk is bad. If the expiration date were closer (lower applicability) or you were more certain the smell was normal (higher applicability), making the applicabilities equal as well, it would be harder to reach a firm conclusion as these confidences would have the same weight. In that case, your tendency would probably be to explore for more evidence, perhaps by cautiously tasting the milk, or, if that were not possible, to choose the conclusion where being wrong would have the lowest cost.

Normal smell BUT you have a cold, 5 days past expiration: Now, because you have a cold, the quality of knowledge conveyed by your own sense of smell is lowered. Therefore the argument that the milk is still good is highly applicable but only probably valid. On the other hand, the evidence that the milk is bad is mostly applicable and based on sound knowledge. Since the validity of the argument against the milk being good is higher, the overall conclusion is that the milk is probably bad. However, Table 1 shows that because the applicability of the argument in favor of the milk being good is higher, you are significantly less confident in that conclusion than you might otherwise have been. In fact, for the overall confidence in the conclusion you subtract a level from the validity of the initial 'against' confidence, leaving you with partly applicable probable confidence that the milk is bad. This significant reduction reflects the conflict and difficulty of deciding for certain between the individual arguments created by these two observations.

These examples demonstrate the importance of both validity and applicability in determining a final confidence. The interplay between these two measures of confidence is complex, but the combination of both elements allows this system to richly express confidence in a conclusion—including handling those situations where the comparative weights of two arguments is not immediately clear. The next

section discusses my implementation of this system of confidence in Calvin, an argumentation system based on LA.

3 OVERALL ARCHITECTURE

Calvin's architecture is drawn primarily from the logical (as opposed to the dialectical) branch of argumentation [Reed and Grasso 2007] and particularly draws on the Logic of Argumentation (LA) of Krause *et al.* [Krause et al., 1995]. In this system, an argument is a tuple containing an assertion and the evidence used to support the assertion. Although the authors of LA discuss the need for a confidence system and present several possibilities, they do not demonstrate any such system on a practical problem.

In a logical argumentation system, arguments are formed, collected, and weighed in a distributed manner. Because of this distributed model, logical argumentation requires fewer rules (because the relationships between different argument chains for and against a premise need not be explicitly defined) and more gracefully handles different classes of reasoning and evidence [Prakken 2005]. This provides significant benefits in building a system to solve real problems. First, building a symbolic system requires the extraction of the rules used for reasoning. The more rules a system requires to adequately solve the problem at hand, the more effort is required to obtain these rules. For many real-world problems, these rules are discovered through lengthy discussions with experts; a significant investment in time and energy. Second, it is quite common for concrete problems to include multiple types of evidence, both qualitative and quantitative, of varying trustworthiness. A basic architecture that can handle both types of evidence gracefully is therefore a superior choice for addressing real problems.

Calvin implements LA using a knowledge base composed of simple rules and an engine for unification of input data and rules. The engine performs this unification using simple backwards chaining. Conceptually, Calvin handles each argument as a collection of trees, functionally equivalent to the heavily-nested tuples used by LA.

Calvin extends LA by defining four specific classes of evidence and a concrete confidence system. Calvin's four classes of evidence are `observations`, simple `calculations`, `simulations`, and `arguments`. `Observations` are direct uses of input data from the user from this particular problem instance. Usually an `observation` is some binary quantity—for example, whether input *x* is less than a threshold value or matches a particular string. Simple `calculations` are generally calculations of simple statistical properties of the input data. A simple `calculation` might find the mean of all *y* entered by the user. Calvin also uses simple `calculations` to find such properties as the maximum or minimum value for a field.

More complex calculations are called `simulations` and usually implement some statistical test on the input data, such as checking for a correlation between two fields or altering the original input data and arguing for a hypothesis with the new input data. `Simulations` are implemented as separate procedures referenced by name in the rule base, allowing them to be as complex as necessary. Allowing the rule base to refer to arbitrary procedures provides a convenient way to add new knowledge to the system, in addition to the usual method of adding new rules to the knowledge base. One drawback of this mechanism is that, although Calvin is designed to allow the easy addition of new rules to accommodate changing scientific fields, adding new `simulations` will require actually writing the code that performs the calculations. However, the gain in power and flexibility for the system is well worth this tradeoff.

Finally, the knowledge Calvin is looking for cannot always be directly gleaned from the input data. In this case it is necessary to build a sub-argument and to use that sub-argument as evidence. These sub-arguments are built to be as complete as possible: sub-conclusions and sub-arguments have the same status in Calvin's engine as top-level arguments.

4 TWO-DIMENSIONAL CONFIDENCE VECTORS IN ACTION

I have repeatedly asserted that this system of confidence assignment and comparison is sufficiently concrete and well-suited to solving real problems. In this section, I discuss two drastically different problems I have applied my confidence system to solving, via Calvin, and the results.

4.1 Cosmogenic Isotope Dating

Cosmogenic isotope dating is a method for discovering how long a landform has existed. Experts take several samples from the landform and calculate the exposure time for each one based on known properties of cosmic rays and how they generate particular nuclides. In theory, all of these samples were suddenly exposed at the time the landform was formed. The expectation is therefore that all samples will be about the age of the landform, perturbed only by small random errors.

However, these exposure times (or 'apparent ages') are rarely all the same. Exposure age measurements for different samples usually differ significantly, sometimes by as much as 10,000 years [Shanahan and Zreda, 2000]. When this happens, the expert must attempt to explain the divergence so that s/he can assign a single age to the landform.

Most explanations for a spread in apparent ages are geologic processes. For example, erosion gradually exposes new surfaces, causing some samples to have exposure ages much younger than the landform age. Samples may have been exposed prior to the creation of the landform, a 'process' called inheritance that results in older apparent ages. Other processes include cover by snow or sand and disturbance of the surface such as by vegetation or animals. Human error in the lab or in the field may also affect apparent ages. As is common in real-world applications, data are noisy and frequently cannot be trusted.

Because they cannot conduct controlled experiments, experts in isotope dating resolve these problems by collecting all the evidence they can find about whether various geologic processes were in operation on the landform. This is an ideal problem for Calvin to address. Experts in isotope dating need a system that can present its reasoning along with its conclusions, handle noisy data and partial support, and deal gracefully with conflicting knowledge and observations.

In fact, Calvin is quite successful at addressing the problem of isotope dating. In a detailed study of its ability to reproduce arguments from 18 randomly selected papers, Calvin was able to produce about 76% of the arguments made by the original authors from the input data. In addition, 99% of the arguments Calvin produced (its precision) were found in the original papers. Even more excitingly, about half of the 'extra' arguments produced by Calvin were determined by a domain expert to actually reveal a significant oversight in the original publication.

In addition, my knowledge acquisition interviews with experts revealed significant parallels between their reasoning methods and Calvin's architecture. For example, multiple experts made statements that supported Calvin's two dimensional confidence representation. Some statements directly supported applicability:

> **Expert:** Hope that if one [sample] doesn't agree it is really obvious, like 16, 16, 17, 15, 100
> **Interviewer:** So you would be more concerned if you had 16, 16, 17, 15, 25?
> **Expert:** Yes

In other words, having a value further from the threshold causes the expert to have more overall confidence in his conclusion. Experts also made statements supporting Calvin's notion of validity:

> **Interviewer:** If you had four samples that agreed and one just a little bit older would it make you feel better if you were worried by the strange one at [sampling] time and what would cause that?
> **Expert:** Well if it were the one sample with different lithology... if the one older one were quartzite [...] it's hard to erode quartz. So are you looking at really different surface erosion issue [...] usually if there is one anomalous age, old or young, you are inclined to think it is an anomaly

In other words, although a sample with an anomalous age is usually an outlier, other factors, such as the lithology of that sample, may overrule that tendency. In the example given by this expert, quartz is

less likely to be eroded, and therefore more likely to represent the true landform age. Finally, some expert statements clearly reveal that confidence has more than a single dimension:

Interviewer: So you had the expected results from most of your samples but as you got to some areas there was inheritance in your samples?
Expert: Yep, or to explain the really young ages one thing is that frost is long believed to be really important in turning over and shattering rocks, so that explains the young ages
Interviewer: So why was previous exposure the better explanation? [...]
Expert: They made sense in a geologic context

That is, although frost is the most likely explanation *a priori* for the spread in apparent ages, the context of the surrounding landforms makes inheritance the overall better explanation.

Calvin's success at solving this process-identification problem in cosmogenic isotope is discussed in more detail in [Rassbach de Vesine 2009].

4.1.1 Opening Bridge Hands

Bridge is a card game where each hand consists of an auction followed by playing the contract set by the auction. Players may choose to pass rather than bidding, and the first person to bid (instead of passing) is said to have 'opened' the bidding. Expert players use a large number of heuristics to decide whether they should open the bidding. These heuristics involve significant give-and-take, as they are frequently contradictory. For example, having sufficient honor cards in a hand is good evidence that one should open the bidding; having honors of low quality implies that one should not.

Calvin is able to implement this complex reasoning using about 25 rules. Unsurprisingly, individual experts in the game frequently disagree about whether to open a particular hand. However, they typically acknowledge the validity of one another's reasoning in coming to their different conclusions. Although I have not yet performed a full study of Calvin's success at solving this problem, I have found that it is able to hold its own in this context: experts presented with Calvin's reasoning agree that its points are valid, even when they disagree with its overall conclusion.

5 CONCLUSION

Argumentation systems are capable of solving real-world problems in a surprisingly human-like way. One challenge of implementing a complete argumentation system that is rarely fully addressed in theoretical systems is the design and implementation of a confidence system. An accurate and rich representation of expert confidence is not immediately apparent. I have designed such a system using two-dimensional confidence vectors and fully implemented it in Calvin, a simple argumentation system for solving real-world problems.

Calvin demonstrates significant success at solving two real-world problems in radically different domains. It uses an identical framework—including confidence representation and combination—to solve both of these problems. Thus, I conclude that the confidence system presented in this paper is broadly applicable to many real world problems that require experts to compare partial, contrasting evidence in order to arrive at a final decision.

ACKNOWLEDGEMENTS

I would like to thank my interview subjects for their generous donation of time and assistance to this project.

REFERENCES

Amgoud, L., Bonnefon, J.F., and Prade, H. 2005. An argumentation-based approach to multiple criteria decision. In Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Berlin, Germany.

Amgoud, L., Cayrol, C., Lagasquie, M.-C., and Livet, P. 2008. On bipolarity in argumentation frameworks. International Journal of Intelligent Systems, 23:1-32.

Cayrol, C., and Lagasquie-Schiex, M.-C. 2003. Gradual acceptability in argumentation systems. In

Proceedings of the Third International workshop on computational models of natural argument: 55-58.

Elvang-Gøransson, M., Krause, P., and Fox, J. 1993. Acceptability of Arguments as “Logical Uncertainty.” In Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Berlin, Germany.

Farley, A.M. 1997. Qualitative Argumentation. In Proceedings of the Eleventh International Workshop on Qualitative Reasoning. Cortona, Italy.

Farreny, H. and Prade, H. 1986. Default and inexact reasoning with possibility degrees. IEEE Transmissions on Systems, Man and Cybernetics 16(2): 270-276.

Krause, P., *et al.*, Ambler, S., Elvang-Gøransson, M., and Fox, J. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence* 11: 113-131, 1995.

Morge, M. and Mancarella, P. 2007. The Hedgehog and the Fox: An Argumentation-Based Decision Support System. In Proceedings of the 4th International Workshop on Argumentation in Multi-Agent Systems.

Prakken, H. 2005. A study of accrual of arguments, with applications to evidential reasoning. In Proceedings of the Tenth International Conference on Artificial Intelligence and Law, 85-94. New York.

Rassbach de Vesine, L. 2009. Calvin: Producing Expert Arguments about Geological History. Doctoral dissertation, University of Colorado. Boulder, Colorado.

Reed, C.A. and Grasso, F. 2007. Recent advances in computational models of argument. International Journal of Intelligent Systems 22(1): 1-15.

Shanahan, T. and Zreda, M. Chronology of Quaternary glaciations on Mount Kenya and Kilimanjaro. *Earth and Planetary Science Letters* 177: 23-42, 2000.

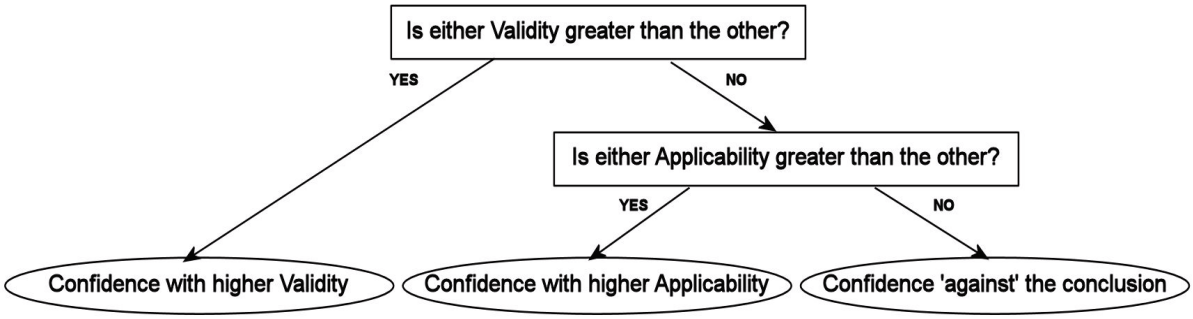

Figure 1: A decision tree for which confidence is considered greater in comparing opposing confidences.

Table 1: Reduction Operations in Confidence Combination

Reduction Operation	Occurs When	
	Validity >	**Validity =**
Do Nothing	Applicability >>	
Applicability - 1	Applicability >=	Applicability >>
Applicability - 2		‘Against’ Applicability >
Validity - 1	Applicability <	
Validity - 1, Applicability - 1		‘For’ Applicability >
Validity - 2, Applicability - 1		Applicability =