

Providing Decision Support for Cosmogenic Isotope Dating

Laura Rassbach, Elizabeth Bradley, and Ken Anderson

■ *Human experts in scientific fields routinely work with evidence that is noisy and untrustworthy, heuristics that are unproven, and possible conclusions that are contradictory. We present a deployed AI system, Calvin, for cosmogenic isotope dating, a domain that is fraught with these difficult issues. Calvin solves these problems using an argumentation framework and a system of confidence that uses two-dimensional vectors to express the quality of heuristics and the applicability of evidence. The arguments it produces are strikingly similar to published expert arguments. Calvin is in daily use by isotope dating experts.*

Automating scientific reasoning is an important challenge to AI. An automated tool can do boring and repetitive reasoning, freeing experts to do more difficult and creative work. Indirectly, it can make explicit the knowledge and reasoning used by experts in the field. Finally, an automated tool can consider all possibilities, sometimes exploring scenarios that human experts may miss.

This article discusses automating reasoning for dating geological landforms. Dating landforms is similar to investigating a crime scene: from the information available on the surface, left behind by an unknown series of events, experts must deduce what happened in the past. In the example diagrammed in figure 1, subsurface rocks are exposed over time as the soil around them erodes. A geoscientist would be faced with the situation shown on the right of the figure; his task is to deduce the situation shown at the left, along with the processes that were at work and the timeline involved.

To accomplish this, a geoscientist first dates a set of rock samples from the present surface, then reasons backward to deduce what process affected the original landform. This is a difficult deduction: geological processes take place over an extremely long period of time, and evidence remaining today is scarce and noisy. Finally, experts in geological dating, like experts in any field, are only human, and can be biased in favor of one theory over another.

In the face of these problems, experts form an exhaustive list of possible hypotheses and consider the evidence for and against each one—much like the AI concept of argumentation. Our system to automate this reasoning, Calvin, uses the same argumentation process as experts, comparing the strength of the evidence for and against a set of hypotheses before coming to a conclusion. We collected knowledge about how isotope dating experts reason through interviews with several dozen geoscientists. Confidence is key in this kind of reasoning, not only in the

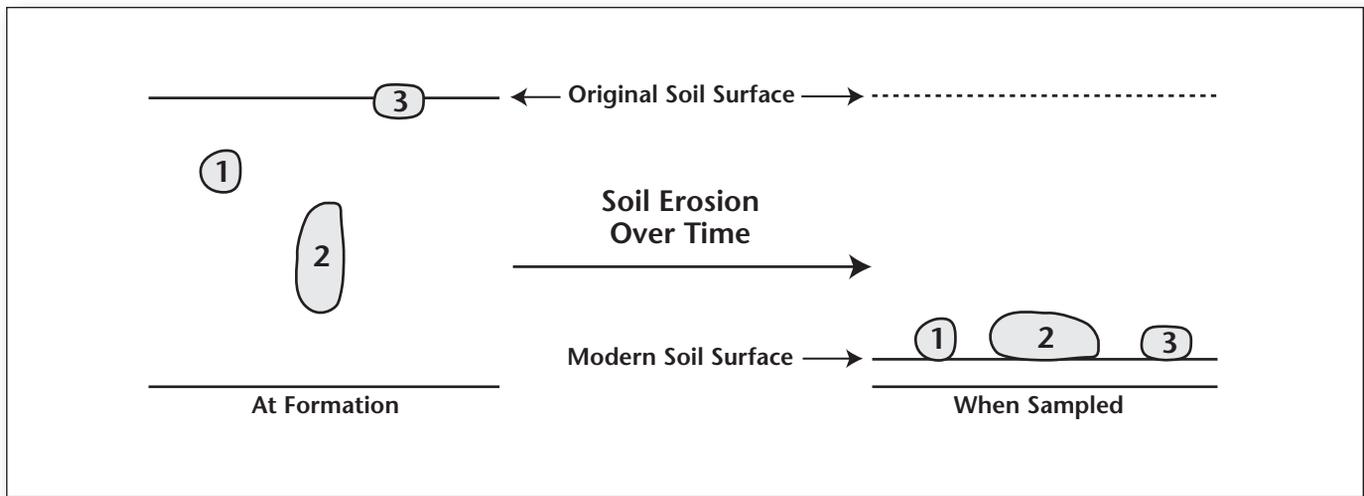


Figure 1. Deducing Past Events from the Evidence Available Now.

quality of evidence, but also in the knowledge that is used to connect evidence to conclusion. Capturing these elements required a novel instantiation of confidence-based reasoning in an argumentation system. From these elements, Calvin produces arguments almost identical to the reasoning presented by human experts.

Calvin provides several contributions to AI and to the larger scientific community. Its rule base is an explicit representation of the knowledge of two dozen experts in landform dating. It incorporates a rich system of confidence that captures the reasoning of real scientists in a useful way. It is a fully implemented and deployed system—a surprisingly rare thing in the argumentation literature. Finally, it is a real tool that is in daily use by real scientists.

In the following section, we discuss the general problem of cosmogenic isotope dating, highlighting its challenges and the approach that experts take to solving it. Next, we describe how Calvin uses argumentation to automate that process, and finally, we discuss our results.

Cosmogenic Isotope Dating

Beginning from a set of samples collected from boulders on a landform, an isotope dating expert's goal is to determine the absolute age of that landform. This section summarizes how experts work, from sampling individual boulders to deducing an age for an entire landform.

The first step is to collect as many samples as possible from the landform. A set of at least five samples is best (Putkonen and Swanson 2003); five to ten samples is about the norm. Experts would prefer to collect far more samples, but often only a handful of boulders suitable for sampling are available. While collecting samples, the expert also

makes qualitative field observations that are often crucial for interpreting initial dating results.

Once the expert has gathered a set of samples in the field, he brings them to a lab for dating. He finds the exposure age of each sample by determining its isotopic composition (some isotopes are produced only by cosmic rays, which do not penetrate soil deeply). Then he performs a series of calculations using this composition and some of the observations taken at the sample site (such as contours of the surrounding area that would impede some cosmic rays, called topographical shielding) to find the length of time the sample has been at the surface. This length of time takes the form of a value with error bars. The expert's next step is to derive an absolute landform age.

For most landforms, the surface exposure times of boulders found on the surface are a true measure of the age of the landform. This is because the boulders are brought to the surface from deep bedrock when the landform is formed. However, different landforms are exposed to different events, complicating the task of determining an overall age for a landform. The simplest version of this problem arises when the expert has a large number of samples and all of their ages overlap, as shown in figure 2a.

Unfortunately, sample sets rarely have a perfect range of overlap. Instead, initial sample ages are usually spread over a wider range than the individual sample errors as in figure 2b. In these cases, the researcher must construct an explanation for the spread in apparent ages, usually a geologic process acting on the samples over time. Once he has found a process that explains the majority of the data, he uses further calculations and educated guesswork to remove its effects from the sample set and (hopefully) arrive at a single age for the land-

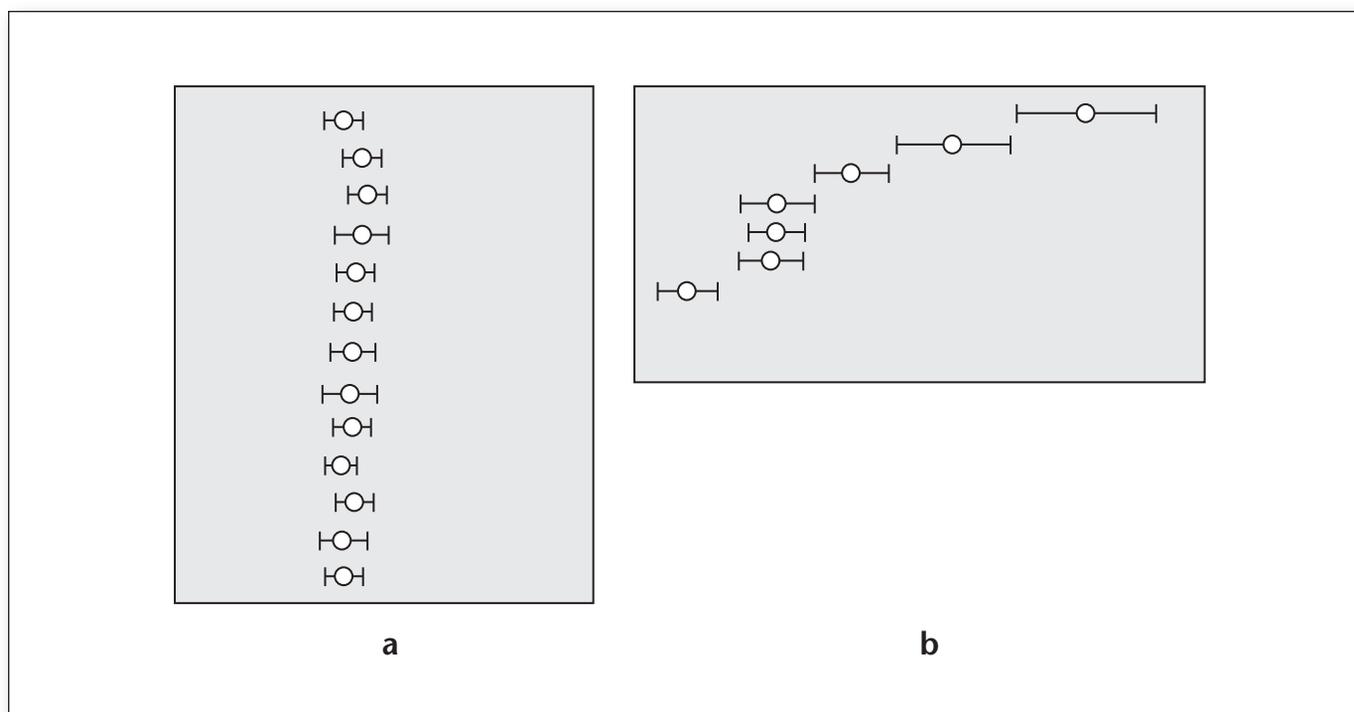


Figure 2. Sample Ages, with Error Bars, Arranged by Sample Age.

a. Ideal distribution. b. Real distribution: Many samples, clustered, small errors; few samples, widely spread, large errors. The ideal distribution is almost never encountered in real cases.

form. In real landforms, more than one process may have been at work, but experts generally focus on isolating the one that most affected the ages of the samples.

Unfortunately, a single round of analysis does not always serve to isolate a landform's true age with any confidence. In this case, the expert must return to the original sample site (at great expense) to seek further samples that disambiguate between possible hypotheses or reinforce the evidence for a likely process. For example, an expensive soil sample at depth can distinguish between several candidate processes.

Most explanations that experts use for a spread in apparent ages come from a short list of geologic processes that affect the exposure times of the samples. Statistical "processes" may also explain the data: for example, the age spread may be a result of lab error or some form of missampling.

Despite the relatively small number of candidate processes, selecting an explanation for the apparent age spread of a particular landform is not a simple task. Available data are noisy and untrustworthy. Experts make mistakes in their observations in the field. Moreover, the manifestation of one process may be quite similar to the manifestations of other processes. Experts usually make a final decision about the process in effect on the basis of heuristic reasoning. These heuristics frequently

contradict each other, and different experts also hold contradictory opinions about the correct heuristics. Addressing this contradiction was a major factor in Calvin's design, as we discuss in the following section.

Design and Architecture

Calvin's input is a set of samples that have already been individually dated (experts use a different tool for this step, such as ACE [Anderson et al. 2007]). It analyzes these groups of samples to determine what process affected the whole landform. Our selection of an appropriate framework for solving this problem rested primarily on data we gathered during extensive interviews. From these interviews, we arrived at an argumentation framework as the best one for Calvin. This selection led to a need to represent both expert knowledge and expert sources and comparison of confidence.

Interviews: Design Motivation

We interviewed 26 experts in isotope dating, amassing around 40 hours of formal interviews and a similar length of informal interviews. Transcripts of formal interviews can be found in Rassbach (2009). We learned that experts in isotope dating consistently use the method of multiple simultaneous hypotheses (Chamberlain 1965).

Then they form arguments for and against every hypothesis, judging their relative and absolute strength to arrive at a solution.

We learned two especially significant things in our interviews. First, experts reason with contradictory heuristics (*inheritance* is the term for samples being exposed before landform formation, making their cosmogenic age older than the landform):

Geologist: The thing about inheritance is, it's usually thought about as quantized, not incremental

Interviewer: So it shouldn't be a spread of ages.

Geologist: Yeah; however, you can convince me you would see a continuum. That is, not only do experts disagree with each other, they sometimes disagree with themselves; and second, experts themselves are convinced that reasoning in their field takes place in the form of argument.

Interviewer: So we're trying to understand what it is that you do.

Geologist: Well, mostly we argue with each other.

The structure of expert reasoning revealed in these interviews makes argumentation a natural framework for automating expert reasoning.

Overall Architecture

Calvin's engine is handwritten for this specific problem. Because the needs for the engine are relatively simple (most of the work of the reasoning is really in confidence combination, described below), writing an engine by hand allowed us to address only our specific domain space. The engine is capable of handling both symbolic reasoning (items that can be essentially true or false) and continuous values. The total engine is about 500 lines of Python code.

At this time, Calvin's knowledge base is composed of 108 rules of varying complexity. The construction of new rules is well-documented in the source code (which is the distribution method), and rules are simple Python structures. The experts we communicated with were generally comfortable writing some amount of mathematical code already. Users are invited to add new rules and it is our expectation they will be able to do so (although we have not been informed of any new rules yet). We are in the very beginning stages of a project allowing users to argue back at Calvin to adjust the confidence in existing rules and add new rules in a more natural fashion.

Reasoning Process

Most processes that affect a landform come from a set list: (1) the possibility that no process at all was at work, (2) exhumation, (3) clast erosion, (4) inheritance, (5) vegetation cover, (6) snow cover, and (7) the possibility that some samples are outliers. Other processes do sometimes affect land-

forms, but these seven are the most common.

Exhumation is samples being exposed (exhumed) after initial landform formation by the erosion of the soil around them (since generally boulders are sampled). Clast erosion is the erosion of the sample boulders themselves. Inheritance, discussed earlier, is prior exposure of sampled boulders. Finally, various kinds of cover can interfere with cosmic rays reaching the surface. Because the possible processes are known, experts do not generally need to form novel hypotheses to find an explanation for their data. Therefore, Calvin gives every hypothesis from the list of "usual suspects" equal consideration, as recommended in Chamberlain (1965) and by experts during our interviews.

Calvin's main task is generating arguments for and against each hypothesis in its list. This process involves finding the applicable information in its knowledge base, unifying that knowledge with the data for the current set of samples, and using that unification to construct a collection of arguments about the conclusion. Performing these functions requires a number of design elements: an engine, rules, evidence, and arguments. Calvin considers candidate hypotheses one at a time and builds arguments for and against each hypothesis from the top down using backward chaining. First, the engine finds all the rules that apply to this hypothesis—that is, those that refer to the same conclusion. Unification is applied to each of these rules, resulting in either a new conclusion to consider or a comparison to input data. Calvin builds the most complete possible set of arguments from its knowledge base for and against each hypothesis. Figure 3 illustrates this backward-chaining process for an argument about the snow cover on a landform. Calvin's engine finds the applicable set of rules, considers each one, and then forms a confidence in the overall evidence. Eventually Calvin will consider every rule about snow cover in its knowledge base and, if the data for unification exists, the rule will be used in its resulting reasoning. Every rule in Calvin contains both a conclusion and a template for evidence that will support that conclusion. The primary portion of a rule is an implication of the form $A \rightarrow C$, where A may be either a single literal or the conjunction (or disjunction) of several literals, and C is the conclusion that A supports. Calvin uses its rules to form an argument (not a proof) for each element in A . From arguments in favor of A , Calvin creates an argument for C . The representation of the argument contains both the rule and the arguments for the antecedents. However, stronger arguments against the conclusion may be found, and Calvin's belief in it overturned. This is the main distinction between an argumentation system and a classical first-order logic system.

Calvin's rules contain several additional ele-

ments that serve important functions: a quality rating, a guard, and a confidence template. The quality rating and confidence template are used to judge the relative and absolute strengths of arguments. Guards prevent the engine from building arguments using rules that are not applicable to the current case. For example, Calvin knows that snow cover is more likely if sample age is inversely correlated with elevation. This is based on the knowledge that snow cover blocks cosmic rays and more snow falls at higher elevations, but only makes sense for sample sets with large elevation ranges. Otherwise, random differences in the data might be interpreted as a meaningful correlation. The guard on this rule tells the engine to ignore the rule unless this precondition holds. Other argumentation systems typically do not require an explicit guard mechanism because they instead defeat rules explicitly (Farley 1997, Morge and Mancarella 2007).

The antecedents in a rule are templates for the evidence that will satisfy that rule, and therefore argue for the rule's conclusion. These patterns define both what evidence is needed to satisfy the rule and where that evidence can be located: usually, the choice is whether to build a subargument for a new conclusion or refer directly to the data input by the user.

The arguments for and against a conclusion *C* are a collection of trees constructed by Calvin's engine by unifying rules with evidence. Alternatively, each argument can be viewed as a tuple of the conclusion and support for the argument, as in the Logic of Argumentation of Krause et al. (1995). The root of each tree in the collection is a rule whose conclusion is *C*, such as the rule $A \Rightarrow C$. Each child of this root is one of the literals in *A* unified with evidence. This evidence may be either additional collections of argument trees or a reference to the input data.

Calvin's backward-chaining engine generally makes no distinction between negative and positive evidence. This is not a valid method in classical logic, where the knowledge that $A \Rightarrow C$ certainly does not imply that $\neg A \Rightarrow \neg C$. However, Calvin's reasoning is intended to mimic that of experts, who are not necessarily logical. Experts not only apply rules in this negative fashion, they regard it as a sufficiently defensible practice that they discuss it in published reasoning. For example, Jackson et al. (1997) include the statement that, since there is no visual evidence of erosion, erosion is unlikely in the area under consideration. Furthermore, the goal of Calvin's reasoning is not to produce logically correct arguments. Rather, it is to produce humanlike arguments. We know that human beings routinely apply rules in this negative fashion (which is why courses in logical thinking put so much effort into telling us not to do so).

Weighing Arguments

Some arguments carry greater weight than others, but precise comparisons between arguments are not always easy to perform. For example, some arguments for exhumation on a hypothetical moraine might be as follows: (1) This moraine has a flat crest, which is a visual sign of matrix erosion. Matrix erosion causes exhumation. (2) This landform is a moraine, and moraines usually have a matrix, which is soft and erodes quickly. Matrix erosion causes exhumation. (3) This landform has samples as old as 50 ky, and various processes often disturb the surface and cause exhumation over such a long time period.

Clearly arguments 1 and 2 are similar, sharing the same root rule. Calvin would derive these arguments as a single tree with two branches. However, experts would consider argument 1 a stronger argument for exhumation because it draws on empirical observations rather than general knowledge about moraines. This issue is often handled in argumentation systems by referring to the specificity of arguments, with more-specific arguments carrying more weight (Elvang-Gøransson, Krause, and Fox 1993). Comparisons are less simple when we consider argument 3: although it refers to information that is specific to this landform, it seems weaker than argument 1. Furthermore, the relationship between arguments 2 and 3 is surprisingly difficult to quantify. How, then, are we to judge the strengths of these three arguments in a way that preserves the intuitive relationships between them?

The central principal of Calvin's confidence system is 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. Defining confidence with two dimensions allows us to clarify why one argument is better than another. Argument 1 uses high-quality evidence and high-quality knowledge. Argument 2 uses high-quality knowledge but only moderate-quality evidence, and argument 3 uses high-quality evidence but low-quality knowledge. Separating the sources of confidence greatly enhances our understanding of the strengths of these three arguments. To instantiate this, Calvin represents confidence as a two-dimensional vector. One element of the vector is determined by which rules were used to form the argument, and the other is determined by how closely the observed situation matched those rules.

As part of our knowledge engineering process, we asked experts about the strength of their belief in their heuristics to determine the appropriate qualitative validity to assign to each rule. Generally, experts were in agreement about what rules were potentially valid and their approximate acceptability. When they were not, our approach

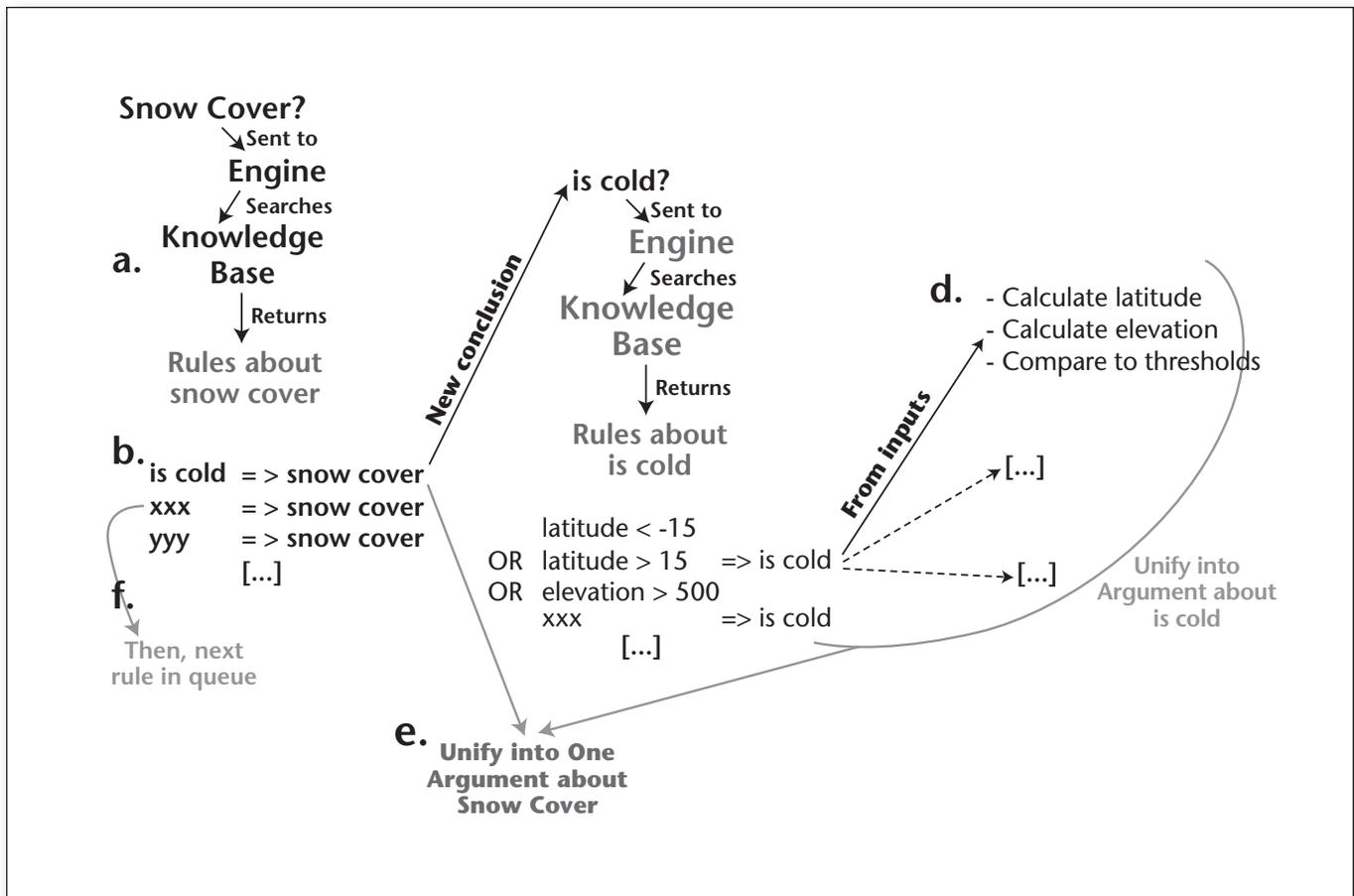


Figure 3. Illustration of Calvin's Chaining of Rules.

(a) Calvin's engine finds all the rules in its knowledge base with a "snow cover" conclusion and puts them in an unordered set to consider one at a time. (b) Calvin considers a rule that snow cover is more likely in cold areas. To apply this rule, Calvin must determine if the sampled area is cold, data not input directly. (c) The main reasoning loop is called with a new conclusion, "the area is cold." (d) Calvin sequentially considers every rule about coldness. We show a rule about coldness that finds the average latitude and maximum elevation of the sample site and compares those values to fixed thresholds. (e) The results of arguing about "is cold" are unified with the original rule about "snow cover." (f) Calvin moves on to the next rule about snow cover.

was to use the lowest confidence expressed (unless this was the position of only one expert, strongly refuted by several others). When Calvin unifies evidence with a rule, it creates a confidence vector for the rule's conclusion from the closeness of the current situation to the rule's thresholds (closer to thresholds gives less confidence) and the validity assigned to the rule. Calvin's engine uses this confidence vector to find an overall confidence in chains of arguments and in sets of argument trees.

Calvin has 12 overall confidence levels, 3 levels of applicability, or how closely evidence matches rules, and 4 levels of validity, or the a priori strength of each individual rule (and by extension, the arguments generated using that rule). Evidence can be *partly* applicable, *mostly* applicable, or *highly* applicable. Knowledge (rules) can be *plausible* (sounds reasonable, but no significant evidence for the theory), *probable* (some evidence, but not established or perhaps an emerging theory), *sound*

(almost all experts accept this theory), or *accepted* (as certain as any scientific theory can ever be). Each confidence level could apply to an argument either for or against a conclusion. Because an intuitive complete ordering for Calvin's confidence levels does not seem apparent, we have imposed the intuitive assumption that validity is more important than applicability, but applicability should certainly have a noticeable effect. This assumption has been the main guiding principle in determining how Calvin works with confidence, especially in difficult situations (for example, when one argument has higher validity but lower applicability than an opposing argument).

Using Confidence. To judge the strengths of the arguments it generates, Calvin manipulates confidence values in two distinct ways. The first operates along a single chain of reasoning: snow cover is more likely in cold areas; this area is cold because

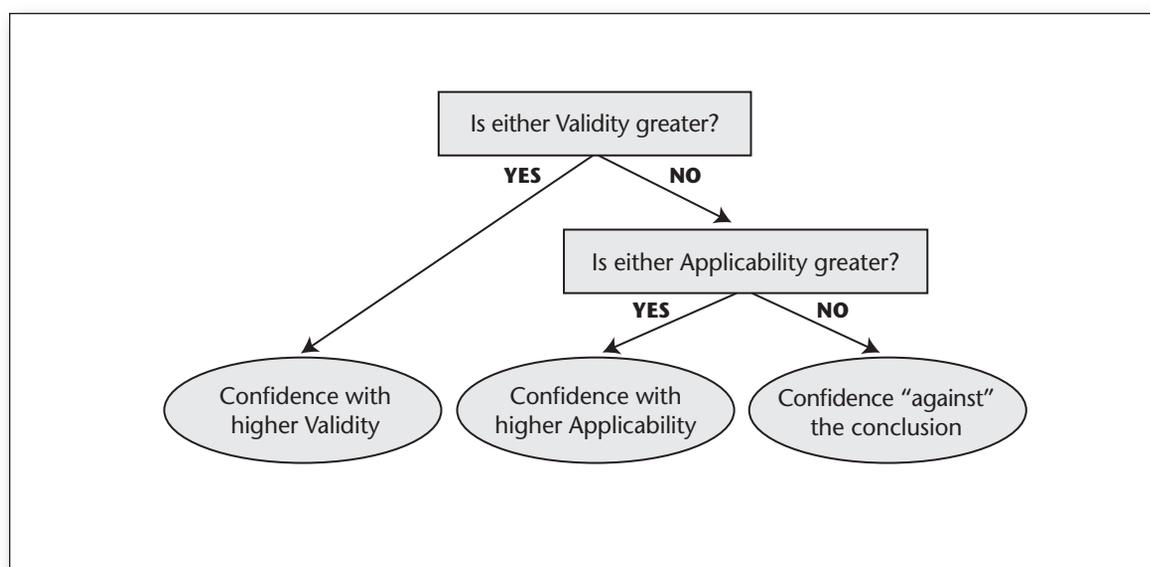


Figure 4. A Decision Tree for Which Confidence Is Greater in Comparing Opposing Confidences.

That is, which argument wins based on having higher confidence.

it is at high elevation. Intuitively, it makes sense to choose the validity of the least-valid rule for the overall conclusion: the chain is only as strong as its weakest link. Applicability is created by the direct use of observed evidence. In this case, how high the sampled area is, compared to what elevation is usually cold, determines the applicability. A few rules 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, for example): (a) erosion is more likely because the landform is old; (b) erosion is less likely because there is no visible sign of it. A chain of reasoning supporting the conclusion might have higher validity but lower applicability than a chain of reasoning refuting the conclusion. There are often several independent chains of reasoning both supporting and refuting the conclusion, each with its own confidence level. Calvin, like many existing argumentation systems (Prakken 2005), assigns confidence in two stages, first locally up a single chain of reasoning and then globally across many chains of reasoning arguing about the same conclusion.

To determine its overall confidence in a conclusion, Calvin first aggregates groups of lower-validity confidences into higher-validity confidences. 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 “frost heaving sometimes occurs in cold areas.” In contrast, a single level of difference in validity is less drastic, for example, the difference between the preceding statement and a statement that “snow cover is plausible in cold areas.” The resulting confidence in other situations is illustrated in figure 4 and table 1. Figure 4 indicates which confidence is considered greater and assigned to the overall conclusion. However, when the two confidences are close, Calvin reduces its overall confidence in the conclusion according to how close the two competing confidences are. Table 1 shows the possible ranks of confidence reduction and when they apply.

An In-Depth Example. Following these operations can be confusing, especially when dealing with a complex and unfamiliar domain. To best illustrate Calvin’s confidence, let us consider a question about whether a particular landform has been subject to matrix erosion. Furthermore, for this illustration, only two observations are available to the expert: the shape of the moraine (whether it is sharply peaked or flat) and the general shape of the sample age distribution. Moraines are generally formed with a sharp crest that flattens over time and expected to have highly clustered sample ages. From these observations the expert can judge the overall validity of whether the landform has been subject to moraine erosion.

Reduction Operation	Occurs when	
	Winner's validity > loser's validity AND	Winner's validity = loser's validity AND
Do nothing	Winner's applicability much greater	
Reduce applicability	Winner's applicability equal or slightly greater	Winner's applicability much greater
Reduce applicability <i>twice</i>		Winner is an argument against a conclusion Winner's applicability is slightly greater
Reduce validity <i>and</i> reduce applicability		Winner is an argument for a conclusion Winner's applicability is slightly greater
Reduce validity <i>twice</i>	Winner's applicability is smaller	Argument applicabilities are equal

Table 1. Reduction Operations in Confidence Combination.

Reduction operations are organized from smallest reduction to largest, from top to bottom.

First, consider the case where you observe that the moraine crest is quite sharp but the sample ages are not perfectly clustered. Calvin would consider the very sharp crest highly applicable evidence (it is strongly observed) and unify it with its knowledge that flat-crested moraines have been subject to matrix erosion. In this case, it draws the (logically incorrect) conclusion that, since this is sharply crested, it has not eroded. The sharpness of the moraine crest is generally a quite accurate indicator of whether it has eroded, so Calvin can draw the conclusion that the moraine has not eroded with high applicability (because the moraine is quite sharply crested) and sound validity (because moraine shape is a good, but not perfect, indicator of a lack of erosion). On the other hand, distributed sample ages are generally also a good indicator that some process (often matrix erosion) has affected the landform: in fact, let us assume that there are enough samples that the indication of a spread in ages is about as valid an indicator as the moraine's shape. However, since the spread is still minimal, the conclusion that the moraine eroded is only partly applicable (because the observation is close to the threshold) with sound validity (equal strength rule). Referencing the figure and table for confidence combination, 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, Calvin selects the "for" confidence. According to the table, 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, Calvin has built a mostly applicable, soundly valid argument that the moraine has not been subject to matrix erosion. This confidence level seems to accurately express the experts' slightly cautious conclusion

(given this example problem in interviews) that matrix erosion does not need to be corrected for.

Now consider another case, where the moraine is not totally flat but not as sharp as before and the sample ages are significantly scattered. In this case, the expert is less confident that the moraine still has its original shape, yielding only mostly applicable sound validity evidence that the matrix has not eroded (same "sound" rule as before, unified with less applicable evidence), and the sample dates are significantly spread (and therefore highly applicable evidence that matrix erosion has taken place). Since the validities of the two confidences are equal, Calvin takes the one with the higher applicability: in this case, the argument for matrix erosion. Referencing the combination table again, we find that for equal validity and a winning applicability one level higher, the actual reduction in overall confidence depends on whether the winning argument is "for" or "against" the conclusion. This is actually intended to reflect cautiousness about the more dangerous conclusion.

For experts in geology, the more dangerous conclusion is the one claiming that a specific process occurred (from one of our interview subjects, "in science, you cannot prove anything, you can only disprove things"). Here, we have an "against" confidence that is one applicability level higher than the defeated "for" confidence, and Calvin reduces the overall final confidence by two levels of applicability, yielding a partly applicable sound argument that the matrix of the moraine has eroded. If the samples were slightly more clustered (lower applicability) or the crest of the moraine were sharper (higher applicability), making the applicabilities equal as well, it would be much harder to reach a firm conclusion as these confidences would have the same weight. In that case, given no further evidence, Calvin would declare that matrix

erosion may have taken place with extremely low confidence. A reasonable user response at that point would be to seek evidence to improve the confidence in some conclusion, perhaps by returning to the sample site and taking a soil sample (or analyzing a previously unexamined soil sample).

Finally, let us consider a slightly more complex case. The moraine has a sharp crest, but there is evidence of soil slides along the side, which could have reshaped a flattened crest. Alternatively, perhaps this expert has routinely been a bad judge of whether a moraine is truly sharply crested. The sample distribution is somewhere between mostly clustered and highly distributed. Now we have reduced the quality of knowledge about whether the matrix has eroded based on the sharp crest (we have reduced the quality assigned to the rule “flat crest \Rightarrow matrix erosion”). The argument that the moraine has not eroded is highly applicable but only probably valid. On the other hand, the evidence that the moraine matrix has eroded is mostly applicable and based on sound knowledge. Since the validity of the argument for matrix erosion is higher, the overall conclusion is that the matrix has probably eroded. However, the combination table shows that because the applicability of the argument against matrix erosion is higher, Calvin is significantly less confident in that conclusion than it might otherwise have been. In fact, for the overall confidence in the conclusion it subtracts a level from the validity of the initial “against” confidence, leaving us with partly applicable probable confidence that the matrix of the moraine has eroded. This significant reduction reflects the conflict and difficulty of deciding for certain between the individual arguments created by these two observations.

Calvin, then, reproduces expert reasoning by considering a set list of hypotheses one at a time, creating arguments for and against each hypothesis. Evidence may take the form of a single comparison or a complete subargument. Calvin then weighs these arguments based on the quality of knowledge and certainty of evidence used to generate them. This weighting

results in both absolute and relative judgments of argument strength, as well as indicating the strongest and weakest points of each argument.

Results

Experts publish some of their qualitative reasoning about a landform when they publish its age. While this presentation is usually incomplete, it typically includes information about both rejected and accepted conclusions. We used these to assess Calvin’s ability to reproduce human expert reasoning. We compared Calvin’s reasoning to the reasoning in 18 randomly selected papers discussing one or more isotope dating problems in detail. These publications provide a broad basis of comparison. To compare Calvin’s output with this prose, we extracted every statement from these papers that made an assertion and distilled it to the conclusion being argued and the evidence presented for that conclusion. We then entered all the data given in the paper, ran Calvin, and compared its output to these argument summaries.

In total, we analyzed 18 papers containing a total of 188 argument/position statements. Calvin performed quite well on arguments published in these papers. It closely reproduced the authors’ arguments 62.7 percent of the time and produced similar arguments a further 26.1 percent of the time. Detailed results are presented in Rassbach (2009). In a few cases, Calvin produced arguments that did not appear at all in the original paper. In one such case, when examining Ballantyne, Stone, and Fifield (1998), Calvin argued that the samples were exhumed. The main evidence for this possibility is a disagreement with ages determined for this landform through other methods. To judge these results, we asked a domain expert to assess Calvin’s new argument. He responded:

I think I see both sides here. From the results, the fact that the ages are younger than the C14 data means that exhumation should be taken very seriously ... there is not much in the way of material that could bury them. However the peaks themselves are eroding

Clearly choosing not to explicitly

address exhumation in Ballantyne, Stone, and Fifield (1998) was a major oversight, given the amount of unclear and conflicting evidence that may or may not be indicative of it. Although Calvin does not give exactly the same argument, it has found a major gap in the reasoning published by these authors.

In some cases, Calvin produced arguments strikingly similar to the statements in the paper. These similarities were especially obvious when the authors of the paper expressed significant doubt about their conclusions. For example, consider this passage from Briner and colleagues (2005):

The ca. 56 ka age on the Jago lateral moraine appears to be a clear outlier that we attribute to inheritance. The age of the Okpilak ridge is uncertain; correlation with the Jago ridge supports the suggestion that the two older boulders from the Okpilak ridge contain inherited isotopes. Alternatively, both ridges might be pre-late Wisconsin in age, and the young age cluster on the Jago ridge records accelerated moraine degradation and consequent boulder exhumation during the late Wisconsin. On the other hand, the stabilization age indicated by the ... ca. 27 ka age is consistent with Hamilton’s (1982) age constraints for deglaciation

This passage refers to a set of three samples on the Jago lateral moraine and four on the Okpilak ridge. On the Jago moraine, two samples are around 27 thousand years old (27 ka), plus or minus about 4000 years, and the third sample is about 56 ka. The authors of the paper argue that this 56 ka sample is an outlier, due to inheritance, based on the sample distribution and independent evidence that the age of this ridge should be around 27 ka. However, a weakness in this argument is that the Okpilak ridge, expected to be about the same age as the Jago moraine, contained two samples at about this 56 ka age. An alternative theory is presented that the older sample on the Jago moraine is the true age, and the younger samples were exposed by matrix erosion more recently.

Calvin finds it quite likely that the 56 ka sample is an outlier and attributes the difference to inheritance. However it, too, grapples with explaining the age of the Okpalik ridge: inher-

itance is supported by correlation with the Jago moraine, the 25 ka expected age, and the climate of the area. However, this implies that two-thirds of the samples from that ridge contain significant inheritance, leading to a conflicted overall argument for inheritance. Calvin also finds significant support for exhumation on both moraines, coming to the same uncertain conclusion as the authors.

Conclusion

Calvin is a fully implemented and deployed argumentation system in use by experts in cosmogenic isotope dating (it has been downloaded 178 times). Thus far, we have not had the resources to study experts' use of the system in detail. However, informal discussions indicate that students are using it to ask more concrete questions about professors' reasoning and some groups are using Calvin to assist in checking that all appropriate hypotheses have been considered.

Because of its nature as a concrete system, building Calvin required us to solve a complex problem: how best to describe and compare confidence. Our solution, a two-element vector to represent confidence, and the associated system for weighing rebutting arguments appears to be novel. This system, while complex in implementation, elegantly captures the argument comparisons we observed experts making.

Calvin is an argumentation system because our goal was to reproduce the structure of expert reasoning. Although isotope dating experts may speak in terms of probabilities and chains of reasoning, they, like most scientists, do not reason in a probabilistically or logically correct manner. Thus, an inflexible system of probabilities or logic would find it difficult to reproduce accurately the reasoning of experts in this field. Expert argument comparisons more closely resemble possibilistic logic (Dubois and Fagier 2005, Farreny and Prade 1996) and our own confidence system than either a Bayesian or pure logic system.

While Calvin's initial results are extremely promising, we are in the process of planning a more rigorous study (with automatic annotation

[White 2009]) to more completely test its success at solving this problem. Furthermore, we believe that Calvin's confidence system will translate well to other problems where weighing competing arguments is difficult—both in other scientific fields such as forensic linguistics and problems in other domains, such as the game of bridge. We hope to identify whether there is a cognitive mechanism that weighs rebutting arguments in a consistent way across domains and, if so, to elucidate that mechanism.

Our hope is that systems that produce humanlike reasoning (as opposed to necessarily correct reasoning) will eventually advance the science of AI in two distinct ways. First, systems that reason like us are more able to reveal both our strengths and our weaknesses, and perhaps someday the fundamental way our minds work. Secondly, we hope that systems that reason in a more human way may have higher adoption rates: people are more comfortable conversing with other people, and groups of people can be more creative in coming up with and reasoning about ideas (though this is of course not always the case). AI systems that reason like people may be able someday to provide this valuable sounding-board function without the cost of training another human domain expert.

References

- Anderson, K.; Bradley, L.; Zreda, M.; Rassbach, L.; Zweck, C.; and Sheehan, E. 2007. ACE: Age Calculation Engine—A Design Environment for Cosmogenic Dating Techniques. In *Proceedings of the 2007 International Conference on Advanced Engineering Computing and Applications in Sciences*. Los Alamitos, CA: IEEE Computer Society.
- Ballantyne, C. K.; Stone, J. O.; and Fifield, L. K. 1998. Cosmogenic ^{10}Be Dating of Post-glacial Landsliding at the Storr, Isle of Skye, Scotland. *The Holocene* 8(3): 347–351.
- Briner, J. P.; Kaufman, D. S.; Manley, W. F.; Finkel, R. C.; and Caffee, M. W. 2005. Cosmogenic Exposure Dating of late Pleistocene Moraine Stabilization in Alaska. *GSA Bulletin* 117(7/8): 1108–1120.
- Chamberlain, T. C. 1965. The Method of Multiple Working Hypotheses. (Reprint of 1890 *Science* article.) *Science* 148(3671): 754–759.
- Dubois, D., and Fagier, H. 2005. On the Qualitative Comparison of Sets of Positive and Negative Affects. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Lecture Notes in Computer Science 3571, 305–316. Berlin: Springer-Verlag.
- 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*. Lecture Notes in Computer Science 747. Berlin: Springer-Verlag.
- Farley, A. M. 1997. Qualitative Argumentation. In *Proceedings of the Eleventh International Workshop on Qualitative Reasoning*. Technical Report, Centre National de la Recherche Scientifique. Paris, France.
- Farreny, H., and Prade, H. 1986. Default and Inexact Reasoning with Possibility Degrees. *IEEE Transactions on Systems, Man and Cybernetics* 16(2): 270–276.
- Jackson Jr., L. E.; Phillips, F. M.; Shimamura, K.; and Little, E. C. 1997. Cosmogenic ^{36}Cl Dating of the Foothills Erratics Train, Alberta, Canada. *Geology* 25(3): 195–198.
- Krause, P.; Ambler, S.; Elvang-Gøransson, M.; and Fox, J. 1995. A Logic of Argumentation for Reasoning Under Uncertainty. *Computational Intelligence* 11(1): 113–131.
- Morge, M., and Mancarella, P. 2007. The Hedgehog and the Fox: An Argumentation-Based Decision Support System. In *Argumentation in Multi-Agent Systems*, Proceedings of the 4th International Workshop. Berlin: Springer.
- 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: Association for Computing Machinery.
- Putkonen, J., and Swanson, T. 2003. Accuracy of Cosmogenic Ages for Moraines. *Quaternary Research* 59(2): 255–261.
- Rassbach, L. 2009. Calvin: Producing Expert Arguments about Geological History. Ph.D. diss., Department of Computer Science, University of Colorado, Boulder, Colorado.
- White, E. 2009. Pattern-Based Recovery of Argumentation from Scientific Text. Ph.D. diss., Department of Computer Science, University of Colorado, Boulder, Colorado.

Laura Rassbach is a senior developer for MapMyFitness.

Elizabeth Bradley is a professor in the Department of Computer Science at the University of Colorado, Boulder.

Kenneth M. Anderson is an associate professor and associate chair of the Department of Computer Science at the University of Colorado, Boulder.