Testing If Reason Score Improves Discourse

Purpose of this Document

This is mainly a reference for frequently asked questions and a place for feedback on all the details of this experiment. It is not meant to build confidence in the results of the experiment as I'm not confident in it.I don't expect anyone to read it though, although several kind people have. Check out the first few paragraphs below and see if you want to keep going or skip to sections of interest.

Goal

One of the many challenges to civil discourse is an inability to reach consensus on basic logical claims (facts). I hope and want to test if we can increase consensus on a subset of those claims that fit a specific logic type. The result may be that overall civil discourse will be less polarized and more successful.

Hypothesis

When people hear a statement claiming something, we naturally evaluate it. Is it accurate? Our amazing brain often does this intuitively based on a combination of past experience and knowledge. No one can see how we came up with that evaluation, even ourselves. When we debate we often compare two evaluations trying to find differences. We seem to intuit that if we evaluated the same information with the same process we would probably agree.

We often give up on finding agreement due to a lack of energy or time as it is too difficult through normal dialogue or prose to fully understand how a claim was evaluated. I want to see if we can increase the chance of agreement on some types of claims by decreasing the work involved in understanding the underlying evaluation process of the claim.

I understand that it will take a long time and trained skills to express the evaluation process for every claim. The hope is that a few people can do the work and many others can gain agreement. Those same trained people can incorporate feedback into the evaluation model to improve agreement.

Limited Claims and Scope

It is not the intent of this experiment to produce a system that can cover all, or even most, types of claims. This exploration of this hypothesis is following the iterative philosophy of Agile software development. We will start with very simple claims and a single simple math model. If this simple model is proven to be useful it can be expanded to include more models. It would be difficult to accurately express the limitations of this first model here so you should know that it is expected to be limited and you can read the document to understand the limits of this first model.

An oversimplified explanation: the evaluation of a claim must be expressible on a scale from absolutely accurate to unknown to absolutely inaccurate. The internal logic is limited to a hierarchy of pros and cons which are averaged together at each level. The leaf nodes of the tree are assumptions and thus start with 100% accuracy and have the same weight as all other assumptions.

Types of logical claims that can be addressed

- yes/no : can be answered with a yes or no. Although in our system it is usually answered with a confidence of Yes.(aka polar, exclusive disjunction)
- Can be answered by weighted reasons and evidence.

Out of scope

- Multiple alternatives or proposals
- Open questions
- Questions with logic outside my current model like one or more sufficient solutions or needing logical “and”,”or”,”not”, etc.

The claims used in this experiment will be chosen based on their ability to be expressed in the current model.

Potential Value of a Score on an Argument Map

We are testing to see if the calculated score reduces the cognitive load of

1. Understanding the results of an argument map
2. Identifying assumptions, evidence, and reasons missing from the map
3. Learning new assumptions, evidence, and reasons
4. Reduce the negative impact of heuristics. Many of these fall under the “Too much information” section such as availability, confirmation, misinformation effect, illusory truth,
5. Reduced rolarity and/or increased agreement

Deferred features

This exploration of this hypothesis is following the iterative philosophy of Agile software development. which emphasizes learning from small experiments on parts of the system as opposed to trying to build a fully functioning theory or system. The following features might combine to provide additional exponential value but are intentionally left out of this experiment. If this experiment indicates the scoring method may have potential the future experiments may include some of these features..

- Complex Decisions: This experiment is focused on simple polar facts that do not have a complex cause-effect web. It can handle complicated but not complex issues as defined by the Cynefin framework.
- Contributions: The mapping process is currently too complex and ill defined to expect the average person to build a useful map. If it can be done by trained mappers then in the future we can experiment to see if it can be modified to be done with less training.
- Distributed Data
- Real-time collaboration
- Claim with multiple sufficient child claims. Right now the calculations assure that if there is a sufficient reason it can be the only one with any importance in that set. When we move to more complex claims we will see if the math needs to be adjusted to account for this situation.
- Explicit types of claims. Many epistemic and causal frameworks categorize the claims into different types. This may be layers onto this model in the future to add additional benefit and analysis.
    - For example the COILS framework has Counterfactuality, Implementation, Linkage, and Significance.
    - The Argument Interchange Format has classifications like Qualifier, Rebuttal, Backing, Warrant
- Issues/ problems and comparing multiple solutions. This mapping does not include a list of the problem or comparing multiple solutions. I hope that this will contribute to that in the future but is considered out of scope for the moment.

Why this experiment may fail

Even with the failures below the lessons learned from this experiment may be of value, but here are the potential failure modes:

- The relation between the top score and the descendant claims may not be clear enough to the casual observer and they do not take the effort to explore the map to find where their reasons differ from the map.
- The cost of building these maps may not be valued highly enough for the perceived improvement in the narrow discourse use cases.
- The simple claims that this solution can address may be too few to provide value in having a process to solve

- Many people will not wish to know the truth and will not engage regardless of the ease.

What is a Reason Map?

A reason map is a term I'm using to describe a way of expressing why we believe what we believe in a way that is easier to understand than written out in prose. In simplest form, it is a main claim with a score that expresses how accurate that claim is based on the evidence and reasons beneath it. It is backed up by a hierarchy of claims where each claim has a score which is either pro or con and if it affects its parents confidence or importance. This allows the reader to quickly understand how each claim impacts the main score and thus the underlying evidence and reasons for the conclusion. To ensure consistency and thoroughness the scores are not arbitrary but follow a predictable set of rules.

Rules of a Reason Map

Maps generally have some rules that organize the information in a way that it can be consumed in parts. For example a traditional map of an area has a rule that information is positioned corresponding to where they exist in physical space. The more northerly an item is, the higher it is on the map. This allows you to look at any part of the map and see items in relation to those rules and thus their physical locations. In the same way a reason map has rules that allow the information to be found based on usefulness to the viewer and explored in a non-linear manner which reduces effort in deriving value from the map.

This starter set of rules and calculations is probably the most simple working model and will not contain many rules necessary to evaluate complex claims. Our preference is to test the simple claims and rules first, If that is beneficial then the rule sets can be expanded and tested for increasingly complex claims.

Main Claim

A main claim is a single assertion that can be scored based on how confident you are in its accuracy. Often called a yes-no or polar question that generally would be true or false. These will be graded in a range though because even though the question should have a yes or no answer we generally don't know the actual answer and we express how confident we are that the correct answer is "Yes". The wording of this statement is important and there are several optimizations I suggest.

Child Claims

A child claim is written in the same way as the main claim but has some additional information that indicates how it is intended to affect our confidence (the score) of this claim's parent..

Pro or Con

The child claim is designated as pro or con to indicate that this claim either increases our confidence in its parent or decreases it. If the claim is written so that some people would see it as a pro and others would see it as a con then it is ambiguous and the impact of the claim should be clearly stated to remove the ambiguity.

Affects Confidence or Importance

Each child claim is designated as affecting its parents' confidence or importance. Our term Importance encompasses what is usually called relevance. If you are asserting the parent claim is irrelevant to the main claim then you are in essence saying it is unimportant. We use the term importance because you can also assert that the claim should have a greater weight than its siblings.

Score Meaning

The score is a range from “highly confident” it is accurate to “unsure” or “unknowable” which is expressed from 100% to 0% respectively. To express that the opposite is true you reverse the wording of the statements and the claim text. To express that you are sure a statement is untrue you would reverse the wording of the statement and adjust the claims appropriately.

100	highly confident is is accurate / assumed / true
75	very confident it is accurate
50	confident it is accurate
25	slightly confident it is accurate
0	unsure of its accuracy or sure it is not determined/ determinable to be accurate

Score Calculation

Rather than arbitrarily choosing a score for a claim, the author adds child pro and con claims to justify the score they wish to express. This makes assumptions and evidence explicit.
Confidence (accuracy)

A claim without children is considered an assumption and is set to 100% confident and 100% importance which makes it the same importance (weight) as every other child of that parent.

A claim with children has a score calculated by doing an average of all the children’s confidence scores weighted by their importance.

$$c_i = \begin{cases} \frac{\sum_{j \in Ch(i)} c_j p_j w_j}{\sum_{j \in Ch(i)} w_j}, & \text{if } Ch.length > \emptyset \\ 1, & \text{otherwise} \end{cases}$$

c	confidence
p	pro/con: 1 for pro -1 for con
w	Importance (weight): see formula below
Ch	Set of child claims that affect confidence

Or, If you want to leave out the defaults and simplify the sum symbol

$$c_i = \frac{\sum_j c_j p_j w_j}{\sum_j w_j}$$

Importance (weight, relevance)

The importance (weight) of a claim decreases as the confidence decreases.The importance can also be affected by child claims that affect importance.A pro child that affects importance takes the importance from above and adds a multiple of 2 times its confidence. This may take it above 100%. Each con child that affects importance adds a multiple of .5 times its confidence to the above importance. This is probably the least thought out part of the formula and may use a different method in the future.

$$w_i = c\left(\begin{cases} \sum_{j \in Ch(i)} c_j m_j, & \text{if } Ch.length > \emptyset \\ 1, & \text{otherwise} \end{cases} \right)$$

c	confidence
w	Importance (weight)
m	importance multiplier. Pro *2 doubles the importance. Con *.5 halves the importance
Ch	Set of child claims that affect importance

Or, If you want to leave out the defaults and simplify the sum symbol

$$w_i = c\left(\sum_j c_j m_j\right)$$

w	Importance (weight)
c	confidence
m	importance multiplier .2 for pro .5 for con
Σ_j	Sum the set of child claims that affect importance

Combined Formula

$$c_i = \frac{\sum_j c_j p_j \left(\sum_{jj} c_{jj} m_{jj}\right)}{\sum_j \left(\sum_{jj} c_{jj} m_{jj}\right)}$$

w	Importance (weight)
c	confidence
p	pro/con: 1 for pro -1 for con
m	importance multiplier .2 for pro .5 for con
Σ_j	Sum the set of child claims that affect confidence
Σ_{jj}	Sum the set of grandchild claims that affect importance

Code for Calculations

You can see the actual JavaScript code that performs these calculations. Please ignore the following variables as they are used for calculating visualizations and probably should be move to a separate function for transparency

The test cases may also help in understanding.

Calculation Notes

The scores are calculated from the leaf nodes to the top ending up with a confidence score (not importance) of the main claim based on the tree of claims below.

In the future there may be exceptions where the author can set the score but the system will have to clearly indicate it was set manually. This calculation was derived by considering how a child might weigh evidence in their mind without any prior experience on the subject.

Example Calculations

[Under Construction]

Wording Guidelines

1. only include a single claim
2. unambiguous
3. Sometimes a claim seems to be both a pro and a con. This is often due to a missing conclusion statement.

Grouping Guidelines

The scoring is affected by the position of the claims in the tree. The positioning of the claims is part of clearly expressing the epistemology of the author. [Needs more explanation]

Example Claims

The claims are phrased as questions for simplicity. To be an actual claim they would be rephrased in the positive or negative. Many of these may be appropriate as the model matures but are out of scope at the moment as I work from simple to complex.

Good Claims for Mapping / Scoring

- “Have concealed carry laws reduced crime?”

Bad Claims for Mapping / Scoring

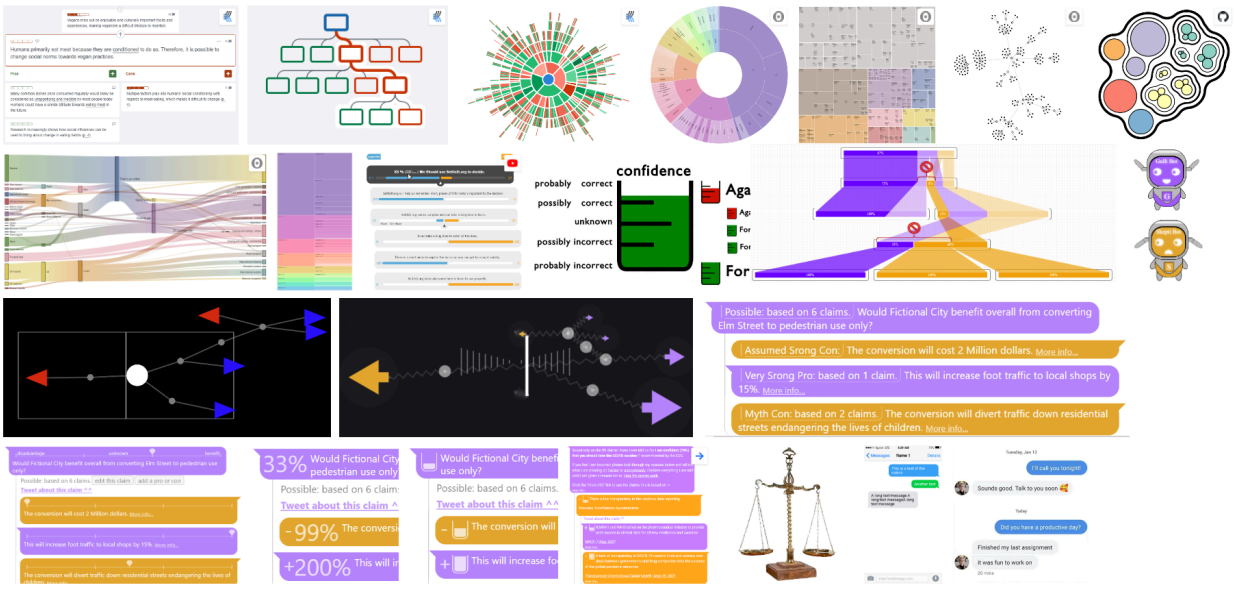
- “Should X state implement a concealed carry law”. The reasons to believe in this claim may be expressible in a reason score map but claims that include “should”

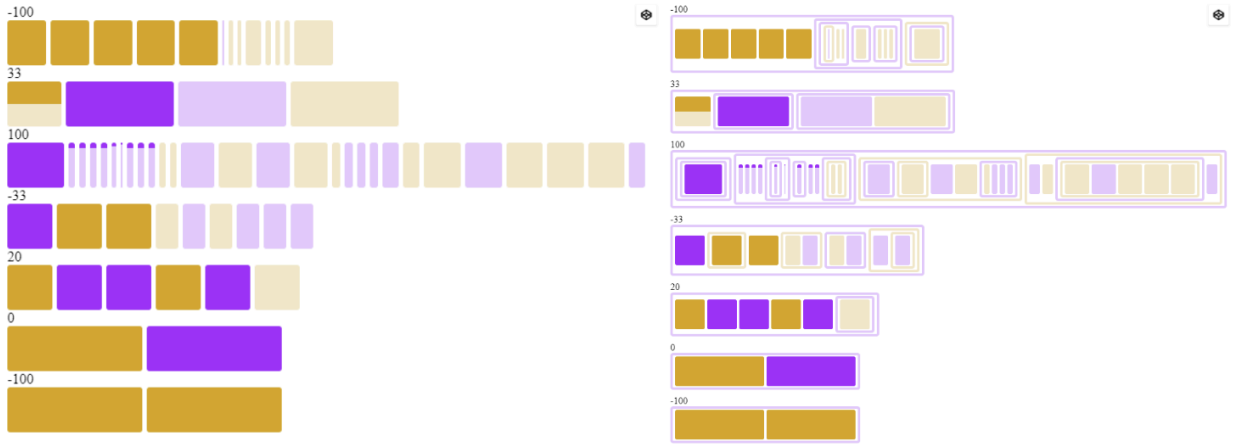
Visualizations

I haven't yet hit on a great visualization. I suspect there needs to be several ways of looking at a reason map and I have several in mind to test out. Here are some ideas and examples.

Bucket Score

Visualization Inspiration





Process of Evaluating Reason Maps

This is still under development. An indication of softening in polarity of an opinion on a claim that an argument map shows is undetermined might be a proxy for improved discourse and increasing openness to other opinions.

1. Recruit people to a study
2. Start with an initial survey question to rate their opinion. For example “Concealed Carry laws reduce crime”
    a. Obviously True
    b. Probably True
    c. Maybe true
    d. Unknown
    e. Maybe false
    f. Probably false
    g. Obviously false
3. Explain how to read a reason score map using the above example
4. Ask them to explore the argument map for a period of time
5. Ask them to provide feedback on the argument map or any child claims in the map.
6. Ask them the initial survey question again
7. Send them a followup email in a week and ask the initial survey question again.

Comparison of Similar Projects and Research

Argument Mapping Research Report - Carefull 2020

> “We found that while argument mapping may be useful in a number of ways, it did not seem promising as a standalone way to obtain substantial (10x-100x) improvements in the ability of groups to reach consensus”

This report outlines the limits of traditional argument mapping that this experiment is attempting to address. I propose that the human brain does not have the ability to calculate the weight of more than 7±2 claims. Humans then have to fall back to heuristics and the difference in heuristics and memory result in different conclusions. By externalizing the simplest calculations humans can work at filling out the map and use the scores to quickly find the differences between their assessments then there is a chance of addressing the differences instead of them being obscured by the number of claims.

> "large groups of unskilled collaborators do not produce quality argument maps"

I agree that at this time the process for creating a map is too complicated and ill defined that it is better to have trained mappers create the maps.Feedback from the public can be incorporated by the trained mappers. This may be a new format for some types of journalism.

Additional note: the example top claims in this study were overly broad and vague and are unsuitable for an argument map and are larger than the scope of this experiment.

Kialo

Kialo has several similarities and several differences. The main difference between Kialo and this project is that scoring in Kialo is based on votes while this score is derived from the position of the claims. Since the Kialo scores do not roll up they can not be used to assess the child claims without reading each one so there is less efficiency in having a score. In this project you can use the scores to cut down on the amount of reading necessary to understand where your understanding differs from the author.

Harrison Durland's Epistemic Maps

After I thought of using the term epistemic map I searched and found Harrison Durland article on "Epistemic maps" for AI Debates? (or for other issues) which had some overlapping goals. The goal of this paper is to test out just the scoring mechanism which is a small subset of Harrison's goals.
Here are the differences using Harrison's 6 key features:

	Harrison's Feature	Bentley's Difference
1	Node and link system with a free/open structure rather than a mandatory origin node or similar hierarchical structure.	I prefer to prioritize the scoring mechanism for individual claims before tackling the bigger question of comparing and contrasting related claims so a single

		root node is sufficient until this smaller model is proven and optimized.This would be a future desire.
2	The links are labeled to describe/codify the relationship, rather than just having generic, unlabeled links that denote "related"	This project only goes so far as to designate children that are pro/con and affect confidence or importance which does not tackle the full scope Harrison is working on. This would be a future desire.
3	An emphasis on "entities" or just "nodes" in a very broad sense that includes claims, observations, datasets, studies, people (e.g., authors), etc. rather than just concepts (as in many mindmaps I've seen)	This project is narrowly focused on claims. Different entities are not necessary to explore the scoring algorithm and the concept so this would be a future concern.
4	Some degree of logic functionality (e.g., basic if/then coding) to automate certain things, such as "Claims X, Y, and Z all depend on assumption W; Study Q found that W is false / assume that W is false; => flag claims X, Y, and Z as 'needs review'; => flag any claims that depend on claims X, Y, or Z..."	This project is only testing one logical functionality. In that case it is better than the editing systems but is not close to Harrison's full vision. I do have in my feature list the ability to have formulas augment the standard formula or bring in outside values but that is after the original concept is proven to be of value.
5	The ability to collaborate or at least share your maps and integrate others'	This is a future goal and was working in past versions of the implementation. I pulled back due to additional complexity and saved it for the future.
6	Preferably some capability for polyadic relationships rather than strictly dyadic relationships	Once the binary accuracy claim is possible then I hope to move to dyadic and polyadic. First I want to solve the simpler case then see if it can be expanded based on those learnings.

History

1980-2010

My family liked debate and we did it often. I collected friends that liked it as well. I was more focused on "winning" by making my opponent give up rather than growing understanding . It ended up being counterproductive in life and I had to unlearn it.

2011 - The Disagreement

In 2011, two of my friends had a disagreement that ended their friendship. I knew I would be seeing each in the next few days so I thought I would do some research on the topic so I could discuss it. It was something that I thought would have been a scientifically answerable question. It was basically "Would it be healthy for most people to eat vegan long term." I was appalled at

how hard it was to find an answer on the internet. Each site felt very biased. It was almost impossible to compare them because their reasons were not clear, worded differently and in different orders and did not address the concerns from the other side.

I started exploring how we can express, map and compare why we believe something. That would hopefully reduce the level of effort required to compare and verify opinions. For the past 10 years I have sporadically experimented with several data models, rules, visualizations and brands to test and refine this idea.

2018 - The Canonical Debate Lab (CDL)

Around 2018 I met several other people working on similar goals of creating a place where people can get all the reasoning about contentious issues from all sides. We vary in our opinions on how best to store and present the information. Reason mapping overlaps with many of the group's ideas but also has some differences. Scoring is not a primary goal of many people in the CDL. You can see more information at CanonicalDebateLab.com.