Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations[⋆]

Markus Brenneis and Martin Mauve

Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
Markus.Brenneis@uni-duesseldorf.de

Abstract. In this paper, we present the results of a study where participants were asked to rate the similarity between sets of positions and arguments. Our goal is to provide a baseline for metrics that compare the attitudes of individual persons in argumentations, with results matching human intuition. Such metrics have different applications, i.a. in recommender systems. We formulated several hypotheses for useful properties, which we then investigated in our survey. As a result, we were able to identify several properties a metric for comparing attitudes in argumentations should have, and got some surprising results we discuss in this paper (e.g., many people do not see a "neutral" position on a line between "pro" and "contra"). For some properties, further research is needed to get a clearer understanding of human intuition.

Keywords: Argumentation · Metric · Human Baseline.

1 Introduction

When discussing with other people, it is interesting to know how similarly another person argues like yourself, i.e. how similar your attitudes are. Do you disagree on central statements, or do you generally agree, but differ in some arguments? Do you have the same priorities for political positions or the same reasons, e.g. for the expansion of wind power? Having a mathematical metric for calculating the (dis-)similarity of attitudes in argumentation enables use-cases like collaborative filtering for argumentation applications like *kialo*[1] or our *deliberate* [5], finding representatives of a group, finding a consensus, and matching political parties and voters based on attitudes and used arguments.

People typically discuss central positions (e.g. the improvement of a course of study [12] or the distribution of funds [8]) and support (or attack) them with other statements, which we call an argument. Each individual person agrees or disagrees more or less strongly with certain statements, and may consider some arguments more important than others when forming an opinion.

When designing a metric for an application where arguments are exchanged, one has to ask which properties that metric should fulfill. For instance, should an

[1] https://kialo.com/

opinion difference in “top-level” arguments against a position (e.g “We should not build nuclear plants, because they are insecure”) weigh more than disagreement on “deeper” arguments (e.g. “Nuclear plants are insecure, because there have been several accidents.”)? Are two persons who are against and for a position equally far apart from each other as two persons where one is for a position, and the other one has a neutral opinion? (Surprisingly for us, our results indicate that the latter is, in fact, the case, as we will explain in Section 4.2.)

Any reasonable metric to answer those questions needs to be based on the perception that humans have regarding the similarity of chains of arguments, instead of the “intuition” of researchers who deal with argumentation theory every day. To establish a baseline for this, we asked our survey participants to judge the similarity of two chains of argumentation. Which pair is considered more similar? The questions asked were based on hypotheses presented in this paper. The hypotheses should help with answering how a metric should behave in trade-off situation, with missing information, hierarchies, and weights in argumentations. To our knowledge, such a survey has not been conducted before.

Our contribution is the following: We formulate several hypotheses for assessing the similarity of argumentations, which should be respected by a metric comparing attitudes expressed in argumentations. We gathered a data set with human assessments of relative similarity of argumentations for testing the real-world relevance of our hypotheses, and checked which hypotheses can be regarded as correct with a high significance.

In the following section, we define central concepts of argumentation theory relevant for this paper. Afterwards, we describe our methods used and our hypotheses. We then present our most important and surprising results. In the fifth section, we discuss our methods, and finally, we comment on related work.

2 Definitions

In this paper, we use terms based on the IBIS model [13] for argumentation. Within an argumentation context, there are *arguments*, which consist of two *statements*: a *premise* and a *conclusion* (e.g., “Nuclear power is sustainable.” can be a premise for “We should build a nuclear power plant.”). When we draw an argumentation graph, statements are nodes, arguments are edges. Statements which are only used as conclusion are called *positions*, and are typically actionable items like “We should build a nuclear power plant”. The unique root of the argumentation graph is called *issue* I, and connects all positions. It is typically the overall topic of the discussion, e.g. “What shall the town spend money for?”.

Each person can have a specific view on the parts of an argumentation graph: A person can agree or disagree with a statement, which we call the person’s *opinion*. Arguments and statements can be of different importance (or relevance, weight) to different persons. Each individual person may use one specific subset of all available arguments. We call the sum of opinions, importances, and arguments used by a person *attitude*.

The results of our work are independent of this model, but it enables us to precisely formulate our hypotheses (i.a. by having statements, not arguments, as atomic elements), and draw graphs for visualizing scenarios for our hypothesis. So our findings can also be applied to metrics working with Dung-style [7] argumentation frameworks; for instance, our issue-based graphs can be transformed to an abstract argumentation framework using the tool *dabasco* [16].

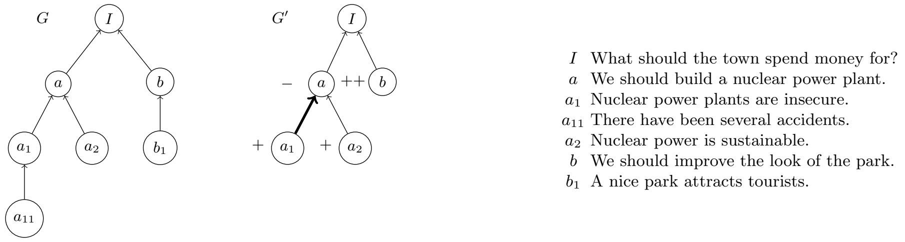

G
I
a
b
a_1
a_2
b_1
a_{11}
G'
I
$-$
a
$++$
b
$+$
a_1
$+$
a_2
I What should the town spend money for?
a We should build a nuclear power plant.
a_1 Nuclear power plants are insecure.
a_{11} There have been several accidents.
a_2 Nuclear power is sustainable.
b We should improve the look of the park.
b_1 A nice park attracts tourists.

Fig. 1: Example for an argumentation graph G and a personal view G' on that graph G with attitudes. Statements with unknown opinion are not drawn in G'.

To understand how our graphs should be read, Figure 1 depicts an example of an argumentation graph G for a discussion and a personal view G' on that graph, which contains Alice's attitudes. In this example, Alice is very sure (++) that she wants the look of the park being improved (b), and she is against a nuclear power plant (a, $-$). She accepts the statements that nuclear power is sustainable (a_2) and nuclear power plants are insecure (a_1), but she thinks the latter weighs more (thick line) for her opinion on building a nuclear power plant. Alice has not mentioned an opinion on the statements a_{11} and b_1.

We will not draw opinions for better readability if the focus of a scenario is not on opinions, and they are considered to be the same across graphs being compared (e.g., "agree"/"+" can be assumed for all statements in Figure 2).

3 Methods

We now present how we developed our hypotheses for properties of a metric for comparing the way different persons or organizations argue, how we created questionnaire scenarios, and conducted the survey. Our focus is explicitly on comparing the *attitudes* of different persons within an argumentation, not properties like number of counterarguments, consistency, or use of rhetorical devices.

We are well aware that our list of properties is only a starting point for the work of finding out how human feeling of argumentation similarity can be translated to a mathematical metric. Thus, we expect that our list can be extended with more properties in the future.

First, we formulate hypotheses about what we expect of a metric. Those hypotheses are at least somewhat reasonable for domain experts, and are partially based on properties of a metric we have presented in an earlier work [4]. However, before they are used for guiding the development of metrics for the comparison of argumentations, it should be checked whether they match the perception of average humans.

To do so, we developed questionnaire scenarios for every single hypothesis. Participants of the survey were asked to assess the similarity of the people's argumentation by indicating which person's argumentation is most similar to the argumentation of another given person. For scenarios which involved only one topic (e.g. an argumentation on nuclear power), we had multiple versions of that scenario with different topics to prevent topic-dependent results.

The survey was conducted using Amazon Mechanical Turk (MTurk) because of its easy and fast recruiting process. Only participants from the US were allowed to assure that there is a sufficient knowledge of English. Although MTurk users are not representative for the US population, it has been shown that the average difference can be quite small [2]. The questions and scenarios were randomly assigned to the participants and the order of answers was randomized. To assure answers of good quality, only answers of participants who answered at least 3 of 5 quality control questions correctly were used in the evaluation.

The complete list of hypotheses is in Table 1. They are grouped in four categories with different motivations: First, we were interested in the influence of basic properties of argumentations, like being for/against a different number of statements and adding arguments. Then we asked ourselves what the influence of weights of opinions and arguments is, and whether they play a role at all. The third group deals with the influence of missing information: Real-world applications often do not have complete information of a person's attitude, how should a metric behave here? The last is about trade-off situations: What weighs more when both, opinions and arguments mentioned, are different between persons? What is the influence if the relevance of positions is rated completely different?

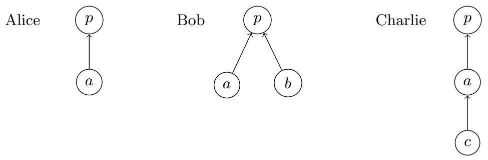


Fig. 2: Visualization of the scenario for Hypothesis 4: The graphs represent the attitudes in the argumentation of each person in the scenario.

As an example, we now present how Hypothesis 4 (*deviations in deeper parts have less contribution to dissimilarity than deviations in higher parts*) has been developed and transformed in a questionnaire scenario. All scenarios can be found in our complete data set which is available online.[2]

[2] https://github.com/hhucn/argumentation-similarity-survey-results

We asked ourselves whether the level where arguments are added is relevant. To make the idea of the hypothesis clearer, Figure 2 depicts the attitudes of the persons involved in the constructed scenario.

Consider Alice, Bob, and Charlie have the same opinions on a position p and a common argument a for it. If Bob adds another argument for p, and Charlie an argument to a, we think that Alice and Charlie are closer because their first-level-argumentation is the same and the deviation is in a deeper part. One could, however, also assume that individuals not familiar with argumentation theory do not have a notion for levels and consider both differences in argumentation behavior as similarly severe.

From our hypotheses, we constructed the following scenario and questions:

> Alice argues as follows on the subject of wind power:
>
> More wind turbines should be built because wind power has a **low environmental impact**.
>
> Bob argues as follows:
>
> More wind turbines should be built because wind power has a **low environmental impact** and because wind turbines are **safe**.
>
> Charlie argues as follows:
>
> More wind turbines should be built because wind power has a **low environmental impact**. The **reason for the low environmental impact** is that they do **not produce any emissions**.
>
> Whose attitude does Alice agree with most?
> – with Bob's attitude – with Charlie's attitude – the attitudes are equally far apart
> Whose attitude does Bob agree with most?
> – with Alice's attitude – with Charlie's attitude – the attitudes are equally far apart
> Whose attitude does Charlie agree with most?
> – with Alice's attitude – with Bob's attitude – the attitudes are equally far apart

The relevant question for us is *Whose attitude does Alice agree with most?* and our expected answer is *with Charlie's attitude*; the other questions were added for gathering additional data and preventing biased answers.

Most other scenarios are constructed the same way. An exception are questions related to missing information, where we asked the questions twice: Once we forced a decision (since a complete, well-defined metric has to make some decision, too), and once we allowed to choose *this cannot be assessed* as an answer.

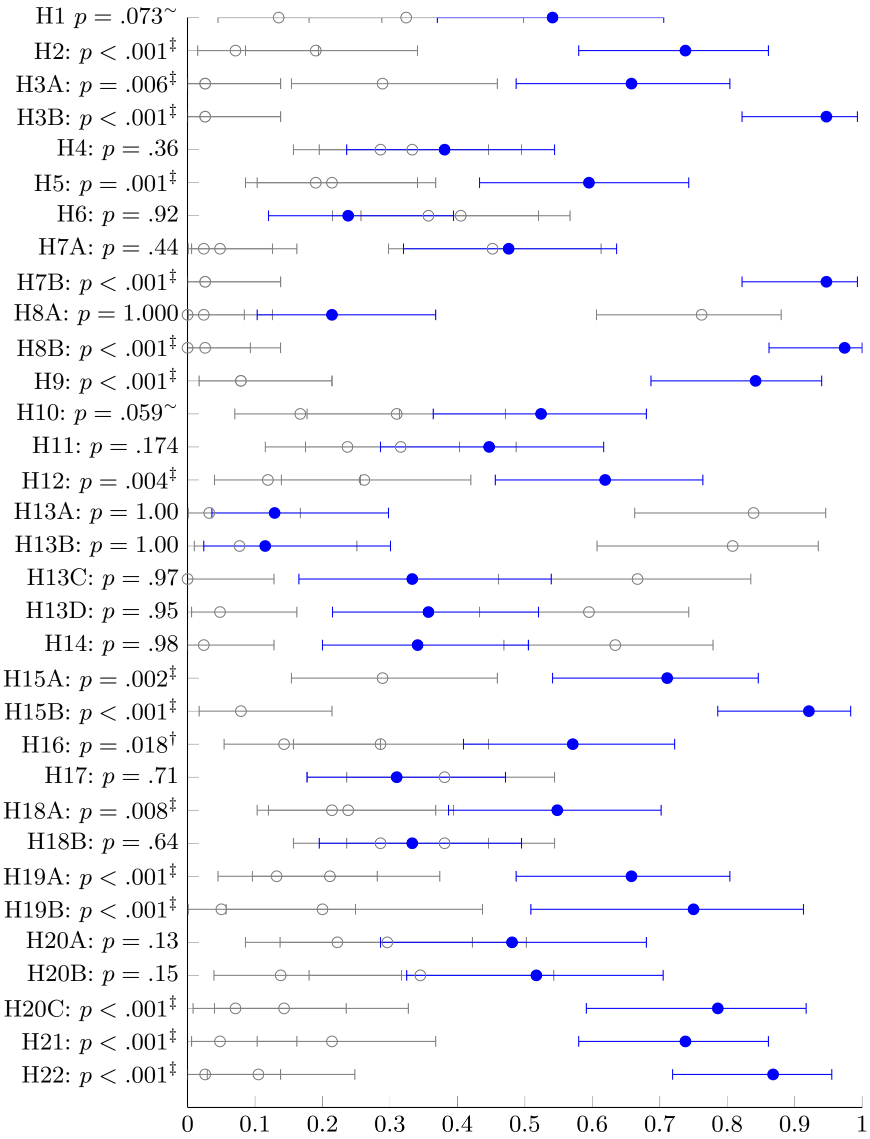
4 Results

We now present the results of our survey, and highlight and explain results which were surprising for us. We report p-values for the null hypothesis "our expected answer is not the most frequently (relative frequency) given answer".[3] For space reasons, not all numbers are presented and discussed in detail, but the aggregated raw data for all questions is available online. A summary of the relative answer frequencies for the relevant questions is depicted in Figure 3.

[3] We used an intersection–union test [18, p. 240] with one-tailed tests on the variances of the difference of two multinomial proportions [9,17], i.e. H_0 is that the differences of the relative answer frequencies between the expected answer and the other answers is not greater than 0.

Table 1: Our hypotheses about the assessment of attitude similarity in argumentations, grouped in basic properties, influence of weights, influence of missing information, and trade-offs

#	Hypothesis
H1	Proportionally bigger overlap of opinions on positions results in greater similarity than the absolute number of differences.
H2	Proportionally bigger overlap on arguments for/against a position results in greater similarity than the absolute number of differences.
H3	A neutral opinion is between a positive and a negative opinion.
H4	Deviations in deeper parts have less contribution to dissimilarity than deviations in higher parts.
H5	Weights of arguments have an influence even if they are the only difference.
H6	Argumentation differences in a branch with lower importance contribute less to dissimilarity.
H7	No opinion is between a positive and a negative opinion.
H8	An unknown opinion is between a positive and a negative opinion.
H9	A statement for which no opinion is mentioned is like a statement for which we explicitly say the opinion is unknown.
H10	Not mentioning an argument and being against an argument have the same effect.
H11	Disagreeing on a position results in greater distance than having the same opinion on that position, but with contrary arguments.
H12	It is possible for a difference in arguments for/against positions to result in greater dissimilarity than a difference in opinions on those positions.
H13	Two argumentations with weak and contrary opinions on a statement can be closer than two argumentations with the same opinions, but with very different strengths.
H14	Two argumentations with weak arguments and contrary opinions on their premises can be closer than two argumentations with the same opinions, but with very different strengths of arguments.
H15	When determining the attitude regarding a position, opinions (not) mentioned for a not-accepted argument have no influence.
H16	Flipping the two most important positions results in a bigger difference than flipping two less important positions.
H17	Adding a new position can remove a previous dissimilarity.
H18	Adding a new position as most important position can swap a previous similarity order.
H19	Agreeing with someone's most important position is as important as having that person's most important opinion matching mine.
H20	Adding another most important position results in greater dissimilarity than flipping the priorities of two positions.
H21	Having more similar priorities of opinions can result in greater similarity even with lower absolute number of same opinions.
H22	Not mentioning a position results in greater dissimilarity than assigning lower priorities.



Clopper–Pearson confidence intervals ($\alpha = 0.05$) indicated with expected answer (blue, filled circles) and other answer options (gray) for the relevant question, p-value for H_0 "expected answer is not the most frequently given answer", ‡: $p \leq 0.01$, †: $p \leq 0.05$, ~: $p \leq 0.10$

Fig. 3: Results for the relevant questions for each hypothesis

After removing participants who did not meet our quality standards, we had, on average, 38 answers for every question relevant for our hypotheses. Those participants have a median age of 30-39 years, which matches the US median of 2018 (36.9). The male/female ratio is 1.96 (total US ratio 0.97), thus we had significantly more male than female participants in our random MTurk sample.

4.1 Results that confirmed our expectations

For many scenarios, we did not get surprising results, and summarize them here.

Proportionally bigger overlap of arguments (H1) or opinions (H2) is indeed more important than the absolute number of differences (H1: expected answer given by 54%, $p = .073$; H2: 74%, $p < .001$). If the assessment of argument relevance is the only difference between attitudes, this is considered as difference by most participants (H5, 60%, $p = .001$).

That the most important opinion in one argumentation matches the opinion in the other argumentation is as important as the reverse case (H19), independent on whether this questions is asked from a person-centric (66%, $p < .001$) or "bird's eye view" (70%, $p < .001$). Flipping the priorities of the most important positions results in a smaller perceived difference than adding a new most important position p, regardless of whether the other persons have not mentioned their opinion on p (H20A, 48%, $p = .13$), had an explicit unknown opinion (H20B, 52%, $p = .15$), or were neutral (H20C, 79%, $p < .001$). Leaving out a position results in a greater dissimilarity than lowering its priority (H22, 87%, $p < .001$). Not only the number of matching opinions on positions is relevant, but, if another argumentation has only a subset of positions, it can be more important that the priorities are more similar (H21, 74%, $p < .001$).

4.2 Surprising Results

We now have a closer look at more surprising findings from survey which were not in line with the expectations we originally had when designing our hypotheses.

No continuum pro–neutral–contra In Hypothesis 3, we conjectured that a neutral opinion lies exactly between a positive and a negative opinion on a statement. As already mentioned in Section 3, we asked this question in two ways: In variant A, "this cannot be assessed" could be chosen by participants, in variant B, a decision has to be made. In both cases, our expected answer ("neutral" is equally far away from "pro" and "contra") was given by most participants (A: 66%, $p = .006$; B: 95%, $p < 0.001$), where the result is much clearer when forced to make a decision.

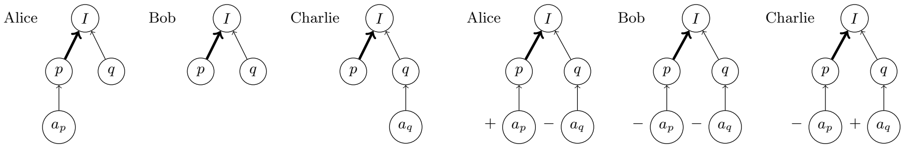
Although the question relevant for us in this scenario was answered as expected, the questions whose attitude is most similar to the positive or negative attitude, respectively, has been answered unexpectedly: We expected that a *positive* opinion is considered closer to *neutral* than to *negative*, but this was only just one of the most frequent answers. In variant B with forced decision, an "equally far apart" assessment has been given by around 50% of the participants.

This can be a hint that many people do not have a mental model where *pro*, *neutral*, and *contra* are arranged in a straight line, but on the corners of a triangle. This might be similar to the opinion triangle presented in [11], with the directions *Belief*, *Disbelief*, and *Ignorance*.

For Hypotheses 7 and 8, we could see similar effects. Hypothesis 7 dealt with whether *no opinion* is equally far away from *pro* and *contra*. For case A, most people give our expected answer (48%, $p = .436$), but many also say that the case cannot be assessed (45%). When forced to make a decision, people choose our expected answer "equally far apart" (95%, $p < 0.001$). But for both variants, we also see the tendency that people have a mental triangle model: In variant B, around 55% have seen *pro* (*contra*) equally far away from *no opinion* and *contra* (*pro*). So being neutral (Hypothesis 3) and having no opinion leads to similar assessments when it is forced, but more people tend to not make an assessment in the *no opinion* case if allowed to.

Lastly, if we consider *pro*, *contra*, and *unknown opinion* (Hypothesis 8), an absolute majority thinks the case cannot be assessed, which makes sense. If a decision is forced, more than 75% percent follow the triangle model again.

Consideration of hierarchies and weights for branches We expected that adding an argument deeper within an argumentation is considered a smaller dissimilarity than adding a new top-level argument (Hypothesis 4, also see Figure 2). This expectation is not confirmed (38%, $p = .36$); the answers are nearly equally distributed across all alternatives. We assume that people count the number of arguments used instead of thinking of an argument hierarchy. Here, further investigations with a more extreme example, e.g. a "deeper" argumentation, would be interesting.



(a) Hypothesis 6

(b) Hypothesis 10

Fig. 4: Visualization of the scenarios for Hypothesis 6 and Hypothesis 10; we expect Bob being closer to Charlie than to Alice in both cases.

Related to this finding are unexpected results for Hypothesis 6: Considering the example depicted in Figure 4a, when comparing Bob with Alice and Charlie, we thought that the similarity to Charlie is greater because the introduced difference is in a branch with lower importance (depicted by a thinner edge). This has not been confirmed, our expected answer is the least frequently chosen answer (24%, $p = .92$). More participants think that Bob is most similar to Alice (40%) or the attitudes are equally far apart (36%).

This is related to the assumption that people do not have a notion for argumentation hierarchy. If people do not catch that a_p and a_q are on the level below p or q, respectively, it makes sense that our expected effect cannot be seen.

But this conjecture is contradicted by the answers for Hypothesis 10, where we thought that not mentioning an argument (as in Figure 4a) and being against an argument (Figure 4b) have the same effect. Our expected answer, Bob is more similar to Charlie than to Alice, is now the most frequently chosen answer (52%, $p = .059$). Thus, our explanations for the unexpected results for Hypothesis 6 do not seem to be correct. Maybe the complexity of the scenario for Hypothesis 10 is so large that people pay closer attention to the nuances of the argumentation. Here, further investigations are necessary.

Trade-off between opinions and arguments Consider a scenario where Alice and Bob have the same opinion on a position, but the arguments are contradictory. Charlie has the same opinions as Alice, but a different opinion on the position. We expect that Alice and Bob are closer than Alice and Charlie (Hypothesis 11) since people probably consider opinions on positions as more important than arguments. Most people answered as we have expected (45%, $p = .174$), but there are also many people saying the attitudes are equally far apart (32%). We can conclude that the common opinion on the position has the greater influence on the assessment of attitude similarity, but arguments also play an important part in the assessment.

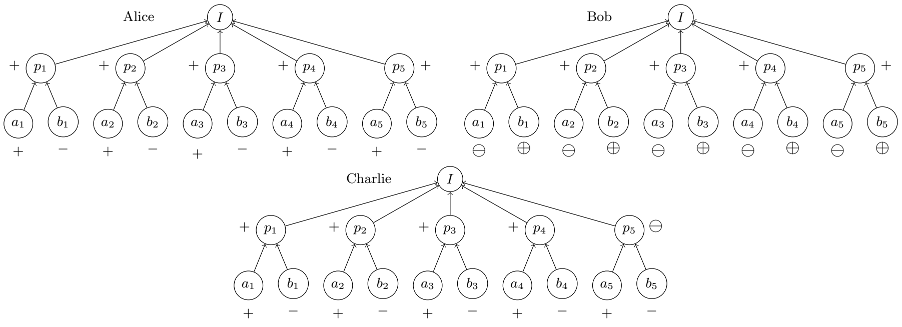


Fig. 5: Visualization of the scenario for Hypothesis 12; differences to Alice are encircled; we expected Alice is considered closer to Charlie than to Bob.

In Hypothesis 12, we assumed not only the opinions on positions are compared, but arguments also play a role and can even "flip" the similarity. For an extreme example with many arguments as shown in Figure 5, our expectation that Alice's attitude is more similar to Charlie's has been confirmed (62%,

p = .004). This is in line with the findings from Hypothesis 11: Not only common opinions on positions are important for assessing similarity, but also the arguments.

Note that our scenario for Hypothesis 12 converges to the scenario for Hypothesis 11 if p_1 to p_4 are removed. As we have only presented those two extreme scenarios in the questionnaire, we cannot say what the "turning point" is, i.e. what number of common arguments is needed to make up for different opinions.

Opinion tendency vs. weight In Hypotheses 13 and 14, we wanted to know whether an argumentation with e.g. weak positive opinion on a position can be closer to a weak negative opinion on the same position than to a very strong positive opinion. We thought that it is possible, but we were proven wrong. The hypotheses were tested with different formulations and scenarios, as strength/weakness can be expressed in different ways: *strongly for* vs. *slight tendency* (A), *for* vs. *no definite opinion* (B), *strongly for* vs. *doesn't really have an opinion* (C), involving a second, common position (D), and *main reason* vs. *very unimportant reason* (E). Our expected answers were not given by most participants (A: 13% p = 1.0; B: 12%, p = 1.0; C: 33%, p = .97; D: 36%, p = .95; E: 34%, p = .98), but the similarity to the person with the same direction of opinion has been rated greater (A: 84%, B: 81%, C: 60%, D: 60%, E: 63%).

We can conclude that opinion tendencies are more important than the weights of opinions and arguments.

Alice argues in favor of wind power as follows:

> I am in favor of wind power, as wind turbines do not produce **CO_2 emissions**. Also, I'm for wind power because **wind turbines look nice**.

Bob argues in favor of wind power as follows:

> I am in favor of wind power, as wind turbines do not produce **CO_2 emissions**. I think **wind turbines look nice**, but that is **no argument for wind power** and not relevant for the discussion.

Charlie argues in favor of wind power as follows:

> I am in favor of wind power, as wind turbines do not produce **CO_2 emissions**. I **don't think that wind turbines look nice**.

Fig. 6: Scenario for Hypothesis 15 on the effect of undercuts: We thought that Bob's and Charlie's attitudes are considered equal.

Understanding of undercuts We expected that an opinion belonging to an undercut argument does not count towards the attitude to a position, i.e. in the scenario described in Figure 6, Charlie's and Bob's attitudes are considered equal, regardless whether Charlie's last sentence is mentioned (case A) or not (B). Our results are not clear for this question: "Do Charlie and Alice [or Bob] have the same attitude (opinion and arguments) on wind power?" has been answered with "Yes" by more than 70% in all cases.

We do not understand this result. It could be that the wording of the question for this case is too technical for a good assessment, so that most people only compared the opinions for the position. Another possible explanation is that untrained persons do not understand the undercut attack correctly or find it confusing, and thus fall back to comparing opinions of positions.

Influence of adding new positions in a priority order We wanted to know how the introduction of a new position by a participant influences similarity order. Our anticipation was that it is possible to remove a previous dissimilarity this way (Hypothesis 17), or even swap the similarity order (Hypothesis 18).

Alice:	Bob:	Charlie:	Charlie':
1. b	1. a	1. a	1. d
2. a	2. c	2. b	2. a
3. c	3. b	3. c	3. b
			4. c

(a) Scenario for Hypothesis 17

Alice:	Bob:	Charlie:	Charlie':
1. a	1. d	1. a	1. e
2. c	2. a	2. b	2. a
3. d	3. b	3. c	3. b
4. b	4. c	4. d	4. c
			5. d

(b) Scenario for Hypothesis 18

Fig. 7: In these scenarios, Charlie' introduces a new position not mentioned by the other participants.

To investigate whether those hypotheses can hold, we checked the scenarios depicted in Figure 7. In Figure 7a, we thought that Charlie is considered more similar to Bob (Hypothesis 16), but Charlie' equally far away from Alice and Bob. The former was confirmed, so changing the order of the most important positions results in a greater perceived difference than flipping less important positions (57%, p = .018). The latter was not confirmed (31%, p = .71), but we see a clear difference from 57%, indicating that the additional position has an influence on the intuition on similarity. There is no clear "correct" answer, though, since the answers are nearly evenly distributed across all alternatives.

For the scenario in Figure 7b, we anticipated that Charlie is closer to Alice (case A), but Charlie' closer to Bob (case B; one way to get to this conclusion is counting the number of absolute place differences for each common statement: Charlie–Alice: 4, Charlie–Bob: 6; Charlie'–Alice: 6, Charlie'–Bob: 4). The first expectation has been confirmed (A: 55%, p = .008), but not the latter (B: 33%, p = .64). In case B, the answers are nearly evenly distributed. Although this is no hint that our hypothesis is sensible, we can see a tendency that the change from case A to B moves the three attitudes closer to each other.

Note that we can neither show that our hypotheses are consistent nor inconsistent, because we only asked for concrete example scenarios. Other scenarios may yield different results, and having results for different scenarios leads to more precise results.

5 Discussion

Our survey was, to our knowledge, the first of its kind. Many results give valuable hints on how an intuitive metric for comparing attitudes expressed in an argumentation should behave. Those metrics have applications in e.g. clustering and recommender systems.

As seen in the previous section, a definite conclusion cannot be drawn for all hypotheses without further surveys. Also, the way we constructed our survey questions could have been suboptimal. We choose a format which is suitable for most Hypothesis to prevent differences due to different formulations of questions. We considered the option to let people rate the similarity of argumentation on a numeric scale, but we thought that this approach is bound to fail: People are unfamiliar with rating argumentation similarity, would probably need some time for "calibration", and the task would feel more unnatural.

Furthermore, the question for "attitude" could have been a problem, because some people may only consider opinions, not arguments. Asking how similar two people "argue" would also be a problem, which we have seen in an internal pretest: Some people started thinking about meta-argumentation aspects, e.g. whether counterarguments are mentioned, or how many arguments are used, and stopped looking at the person's actual attitude.

For questions with ratings of several positions, we switched between complete sentences and enumerations, depending on the number of positions. We thought complete sentences with many positions distract from the actual differences. The change of format could, of course, have an influence, which we did not measure.

We are well aware that MTurk workers are not a representative sample of the US population, and even less for other countries; as already mentioned, the gender distribution does not match the US population. Therefore, generalizing our results for other populations is only possible with caution. Nevertheless, we get some useful insights and hints for further, representative, bigger studies, and possible comparisons between different populations.

6 Related Work

We know no other surveys on attitude similarity in argumentation, but there have been surveys for other purposes to find human baselines.

[14] proposes different measures for determining the similarity of words, and compares the measures with human ratings from a dataset created by [15]. They also think that the quality of a metric can best be determined by comparing it with human common sense. Their dataset contains absolute ratings from 0 (no similarity) to 4 (synonym) for 30 word pairs, each assessed by 38 subjects. We do not think that an absolute rating would have worked for our experiment. First, our argumentation scenarios can have fine-grained or large differences, which probably makes it hard for a person without argumentation theory background to map the difference on a small absolute scale. Second, an absolute scale works well when you can grasp every pair to compare at once and correct older decision

to tweak one's brain scale; this works well with short word pairs, but not with more complex descriptions of argumentation.

In the context of word similarity, [6] find that "comparison with human judgments is the ideal way to evaluate a measure of similarity", which supports our initial assumption that gathering human judgments is important.

In [3], which is based on the study design of [15], 50 human subjects assessed the similarity of process descriptions on a scale from 1 to 5. They compared those assessments with the values of five metrics. Each subject had to indicate how they come to their decision for each comparison, by letting them choose a strategy (e.g. "by process description") from a menu. We did not ask participants how they have come to their assessments. Firstly, we think that reflecting on one's decision influences further decisions. We also think that writing an own description of the decision process is too hard, and providing a menu with possible answers could have influenced following decisions. Moreover, asking this for every question would have significantly increased the length of the questionnaire.

Metrics and applications for comparing argumentations already exist, e.g. based on cosine similarity for opinion prediction [1], and for comparing one's own argumentation with others by counting the number of agreements/disagreements on statements [10]. In both cases, no justification is given why the similarity measure is a good choice. With our work, we want to fill that gap. For instance, we showed that simply counting agreements is not enough.

7 Conclusion and Future Work

We have conducted a survey with human subjects who had to assess the attitude similarity of argumentations. Our results are available for download, and can be used as basis when developing a metric for measuring attitude similarity in argumentation-based applications, e.g. for collaborative filtering. Our results help to transform human gut feeling into a mathematical metric. Some intuitive hypotheses were confirmed by our results, but there were also surprising results, e.g. *neutral* is often not seen as falling on a line between *pro* and *con*.

Our survey cannot establish "absolute truths", but we have collected first hints on what properties a metric which matches human intuition should have. In future work, we want to compare several metrics to see which properties they fulfill and how that matches human intuition. Moreover, further research is needed for hypotheses where we could not get clear results, and where there are turning points in trade-off scenarios. Also, more representative surveys and a comparison of different countries are needed.

References

1. Althuniyan, N., Sirrianni, J.W., Rahman, M.M., Liu, X.F.: Design of mobile service of intelligent large-scale cyber argumentation for analysis and prediction of collective opinions. In: International Conference on AI and Mobile Services. pp. 135–149. Springer (2019)

2. Berinsky, A.J., Huber, G.A., Lenz, G.S.: Using mechanical turk as a subject recruitment tool for experimental research (2011)
3. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? towards personalized similarity measures in ontologies. In: Wirtschaftsinformatik 2005, pp. 1347–1366. Springer (2005)
4. Brenneis, M., Behrendt, M., Harmeling, S., Mauve, M.: How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In: Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020). pp. 2–13. No. 2672 in CEUR Workshop Proceedings, Aachen (Sep 2020)
5. Brenneis, M., Mauve, M.: deliberate – Online Argumentation with Collaborative Filtering. In: Computational Models of Argument. vol. 326, p. 453–454. IOS Press (Sep 2020)
6. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and other lexical resources. vol. 2, pp. 2–2 (2001)
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
8. Ebbinghaus, B., Mauve, M.: decide: Supporting Participatory Budgeting with Online Argumentation. In: Computational Models of Argument. Proceedings of COMMA 2020. Frontiers in Artificial Intelligence and Applications, vol. 326, p. 463–464. IOS Press (Sep 2020)
9. Franklin, C.H.: The 'margin of error' for differences in polls. See https://abcnews.go.com/images/PollingUnit/MOEFranklin.pdf (2007)
10. Gordon, T.F.: Structured consultation with argument graphs. From Knowledge Representation to Argumentation in AI. A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday pp. 115–133 (2013)
11. Haenni, R.: Probabilistic argumentation. Journal of Applied Logic **7**(2), 155–176 (2009)
12. Krauthoff, T., Meter, C., Mauve, M.: Dialog-Based Online Argumentation: Findings from a Field Experiment. In: Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence. pp. 85–99 (November 2017)
13. Kunz, W., Rittel, H.W.J.: Issues as elements of information systems, vol. 131. Citeseer (1970)
14. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering **15**(4), 871–882 (2003)
15. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and cognitive processes **6**(1), 1–28 (1991)
16. Neugebauer, D.: DABASCO: Generating AF, ADF, and ASPIC+ Instances from Real-World Discussions. In: Computational Models of Argument. Proceedings of COMMA 2018. pp. 469–470 (2018)
17. Scott, A.J., Seber, G.A.: Difference of proportions from the same survey. The American Statistician **37**(4a), 319–320 (1983)
18. Silvapulle, M.J., Sen, P.K.: Constrained Statistical Inference: Order, Inequality, and Shape Constraints, vol. 912. John Wiley & Sons (2011)