Related Explanations in Formal Argumentation, an Empirical Study

Roos Scheffers[1], Floris Bex[1,2], and AnneMarie Borg[1]

[1]Department of Information and Computing Sciences, Utrecht University
[2]Tilburg Institute for Law, Technology and Society, Tilburg University

Abstract

In formal argumentation theory, multiple argumentation-based explanation methods have been formulated based on ideas from social and cognitive science. However, these have not yet been empirically validated. One such idea is that information in an explanation needs to be related; in argumentation-based explanations, this has been captured as there being an attack path between arguments. This study describes and empirically validates two types of relatedness, related admissibility and directly related admissibility. This was done by instructing participants to select arguments from an argumentation framework to explain another argument in this framework. These explanations selected by the participants were compared to argumentation-based explanations that use relatedness. We found that both forms of relatedness are cognitively plausible. This gives insight into how argumentation theory can be applied in the real world to provide explanations.

1 Introduction

Abstract argumentation frameworks (AFs) [8] present dialectical argumentation as arguments and the attack relations between them. Given an argumentation framework, semantics are used to determine extensions, sets of arguments which are acceptable together. However, such sets can be very large, and a user might only be interested in the reasoning behind the acceptability of a single argument. Therefore, several argumentation-based explanation methods have been defined [3, 7, 9, 12], which select some arguments in an AF to explain the (non-)acceptance arguments. These explanation methods are motivated by research in the social and cognitive sciences which, according to [13], argues that explanations should be selective. All of these argumentation-based explanation methods use *relatedness* for argument selection: arguments only have explanatory value if they are related to the argument to be explained via attack relations (i.e., they attack or defend the argument). However, whether these explanation methods reflect human cognition, making them cognitively plausible, has not been tested.

The exploration of cognitive plausibility in argumentation began with the work of Rahwan et al. [16]. They advocated for the empirical evaluation of argumentation theories as an essential addition to the traditional example and principle-based evaluation methods. Others have followed Rahwan and empirically investigated other elements of argumentation, such as types of attacks [5], semantics [10], and support [14]. However, to our knowledge, no empirical study has investigated argumentation-based explanations. This gap in the literature is problematic because explanations are fundamentally cognitive acts in which people adjust the information they present based on the audience and context [13]. Testing theories of explanation in context with people is the only way to determine whether they are acceptable to people. Theoretical definitions of explanations without an empirical foundation lack this real-world context.

In this paper, we are the first to experimentally investigate the relatedness of arguments in argumentation-based explanations, thus testing the cognitive plausibility of such explanations. We will test two hypotheses based on relatedness where the second is more restrictive than the first. These hypotheses are: *people*

The final publication is available at IOS Press via https://doi.org/10.3233/FAIA240327.

include only related arguments in explanations and *people include only directly related arguments in explanations*.

The two hypotheses were tested using an online study in which participants were presented with four or five arguments and tasked with explaining the acceptance of one argument by choosing from other arguments. These arguments were presented in natural language and created to instantiate two AFs, a process that was validated using a pilot study. The explanations chosen by participants showed that explanations based on relatedness are indeed cognitively plausible since participants' choices of explanations aligned with both tested forms of relatedness.

In the next section, we give a brief overview of related work. Then, we introduce relevant definitions and the two hypotheses of this paper (Section 3). In Section 4, we describe the pilot and empirical study conducted to test these hypotheses. The results of this study are described in Section 5. Finally, we discuss conclusions and suggestions for future empirical work into argumentation-based explanations.

2 Related work

As far as we are aware, no study has empirically investigated specifically argumentation-based explanations [9, 3, 7, 12]. Therefore, this section provides a general overview of empirical cognitive studies in argumentation. The presented studies are most related to the current work because they investigate the relationship between human cognition and formal models of argumentation. A complete review can be found in [4].

Rahwan et al. [16] conducted the first experiment validating argumentation theory using human cognition. They tested the cognitive plausibility of simple and floating reinstatement. Participants were presented with natural language arguments based on an argumentation framework and then asked to rate the acceptability of the conclusions of those arguments. The ratings provided by participants were compared with the acceptability of arguments according to argumentation semantics. Bezou Vrakatseli et al. [1] replicated the experiment conducted by Rahwan et al. and tested different methods of presenting arguments. They found that seeing all possible scenarios before rating arguments increases participants' acceptability ratings of those arguments.

Guillaume et al. [10] continued this line of research by empirically studying which semantics best align with human judgments of the acceptance of arguments, the existence of attacks, and the directionality of attacks. To validate the natural language argument sets, they had participants draw all attack relations between arguments in a set, and then they compared these to the intended attacks. They found that participants' judgements on the existence and direction of attacks align with definitions in formal argumentation.

Rosenfeld and Kraus [17] predicted argumentative behaviour using argumentation theory, machine learning, and relevance heuristics. These methods were used to predict the next argument to be used in a conversation. They found that the relevance heuristic based on path length is the best predictive metric for this task. Empirical work has also been done on model values [2], concern [11], and persuasion [14] in argumentation. However, none of these studies investigate explanations using (related) arguments in abstract argumentation frameworks.

3 Formal preliminaries and hypotheses

The purpose of this section is to provide definitions of fundamental concepts so that the hypotheses for the study can be formulated at the end of the section. This will include Dung's [8] abstract argumentation frameworks, admissibility, and two methods for selecting related argumentation-based explanations.

An *abstract argumentation framework* (AF) [8] is a pair $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments, and $\mathcal{R}$ is a binary attack relation on these arguments. For every $(A,B) \in \mathcal{R}$, A attacks B. AFs can be represented as a directed graph, where the nodes represent arguments and the arrows represent the attack relations (Figure 1). Given an AF, sets of arguments that are collectively *acceptable*, can be determined [8].

Definition 3.1. Let $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$ and, $A \in \mathcal{A}$, and a set of arguments $S \subseteq \mathcal{A}$. Then:

- A is *acceptable* with respect to S iff each argument attacking A is attacked by any $B \in S$;

- S is *conflict-free* iff there exists no $(A,B) \in \mathcal{R}$ with $A,B \in S$; and
- S is *admissible* iff its is conflict-free and every $A \in \mathcal{A}$ is acceptable w.r.t S.

Admissibility serves as a criterion to identify sets that can be collectively accepted because they hold up against counterarguments. An admissible set can function as an explanation for what Fan and Toni [9] call a topic argument, an individual argument that is to be explained. However, because admissibility does not define how elements of an admissible set are related to each other, it can include numerous arguments that are not related to the topic. These arguments hence do not have explanatory value for the topic argument. In this paper, we consider two argumentation-based explanation methods that consider the relation between a topic argument and the arguments in its explanation: *related admissible explanations* and *directly related admissible explanations*. The below definitions are adapted from [3, 9].

Definition 3.2 (Defends)**.** Let $\mathcal{AF} = (\mathcal{A},\mathcal{R})$ and $A,B,C \in \mathcal{A}$. An argument B *defends* an argument A iff:

1. B attacks C and C attacks A (*direct defense*); or
2. B defends C and C defends A (*indirect defense*).

This defends relation can be combined with admissibility to define related admissible explanations.

Definition 3.3 (Related admissible explanation)**.** Let $\mathcal{AF} = (\mathcal{A},\mathcal{R})$, $A \in \mathcal{A}$ and $S \subseteq \mathcal{A}$. S is a *related admissible explanation* for topic argument A iff A is defended by every argument in S and $S \cup A$ is admissible.

A related admissible explanation has a topic argument that is defended by every argument in the explanation, that is, there will be a path between the topic argument and every other argument in the related admissible set. Note that Fan and Toni [9] define a related admissible set (essentially an extension) in which the topic argument A is also included. However, we decided to not include the topic argument and focus on the explanation.

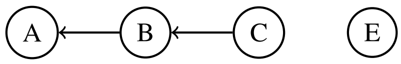


Figure 1: Argumentation framework $\mathcal{AF}_1$.

Example 3.1. Figure 1 shows $\mathcal{AF}_1 = (\mathcal{A}_1,\mathcal{R}_1)$ where $\mathcal{A}_1 = \{A,B,C,E\}$ and $\mathcal{R}_1 = \{(B,A),(C,B)\}$. A possible instantiation of this AF, following conventions set by [16] and [10], is:

(A) Johan got a passing grade on the exam and handed in the course assignment on time, so Johan will pass his course.

(B) Johan handed in the course assignment past the deadline, so he did not hand in the assignments on time.

(C) Johan got an extension for the assignment, so he did not hand it in past the deadline.

(E) Johan has missed two lectures, so he does not have perfect attendance.

Take argument A as the topic argument. Argument E, while in an admissible set together with A, is unrelated to A since there are no attacks to or from E. Therefore, the related admissible explanation for topic argument A would be $\{C\}$.

The second form of relatedness concerns directly related admissible explanations, following Borg and Bex's [3] idea of including only direct defenders in an explanation. Where standard related admissible explanations only consider whether there is a connection to the topic argument, here we will only consider arguments directly related to the topic argument using direct defenders.

Definition 3.4 (Directly related admissible explanation)**.** Let $\mathcal{AF} = (\mathcal{A},\mathcal{R})$, $A \in \mathcal{A}$, and $S \subseteq \mathcal{A}$ be a related admissible explanation according to Definition 3.3. S' is a *directly related admissible explanation* for topic argument A iff $S' \subseteq S$ and A is directly defended by every argument in S'.

A direct defender has a closer connection to the topic argument and is considered to be more closely related than any non-direct defender. This is supported by the relevance heuristic found by Rosenfeld and Kraus [17], which states that an argument is more relevant to another and preferred by people if the path length between them is shorter. Since the path length to a direct defender is always shorter than to an indirect one, this heuristic suggests that people would prefer direct related admissible explanations over standard related admissible explanations.

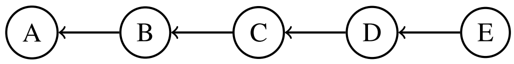


Figure 2: Argumentation framework $\mathcal{AF}_2$

Example 3.2. Figure 2 shows $\mathcal{AF}_2 = (\mathcal{A}_2, \mathcal{R}_2)$ where $\mathcal{A}_2 = \{A, B, C, D, E\}$ and $\mathcal{R}_2 = \{(B,A), (C,B), (D,C), (E,D)\}$. A possible instantiation of this AF is:

(A) The lake is frozen so Dana can go ice skating tomorrow.

(B) According to the weather broadcast the temperature is going to rise, so the lake won't be frozen tomorrow.

(C) The temperature broadcast is unreliable, thus temperature might not rise.

(D) This broadcast is part of a national news show, thus it is not unreliable.

(E) The national news show and the weather broadcast have different funders and staff, thus the broadcast is not part of the news show.

If argument A is the topic argument for an explanation, the related admissible explanation is $\{C, E\}$. The *directly* related admissible explanation for A is $\{C\}$, consisting only of direct defender C.

Two approaches to selecting related arguments for argumentation-based explanation have been described: related admissible and directly related admissible explanations. The latter is more selective than the former. We consider both in this paper since it is not yet known how selective people are when selecting related information to include in explanations based on formal argumentation, therefore both methods should be tested. Therefore, we will consider two hypotheses based on the two forms of relatedness.

H1 *People include only related arguments in explanations.* According to this hypothesis, people adhere to related admissible semantics and, as a result, only include arguments that are connected to the topic argument in their explanations.

H2 *People include only directly related arguments in explanations.* This hypothesis proposes that people follow directly related admissible semantics and, therefore, favour arguments closer to the topic argument.

4 Method

The two hypotheses of this study were tested using explanations provided by participants for arguments in natural language argument sets corresponding to $\mathcal{AF}_1$ (Figure 1) for *H1* and $\mathcal{AF}_2$ (Figure 2) for *H2*.[1] This section provides an overview of the data collection and analysis procedures. In this section, we will outline the study design, procedure, and analysis steps, as depicted in Figure 3.

[1] https://www.uu.nl/en/research/ai-labs/national-policelab-ai/related-explanations-in-formal-argumentation-an-empirical-study

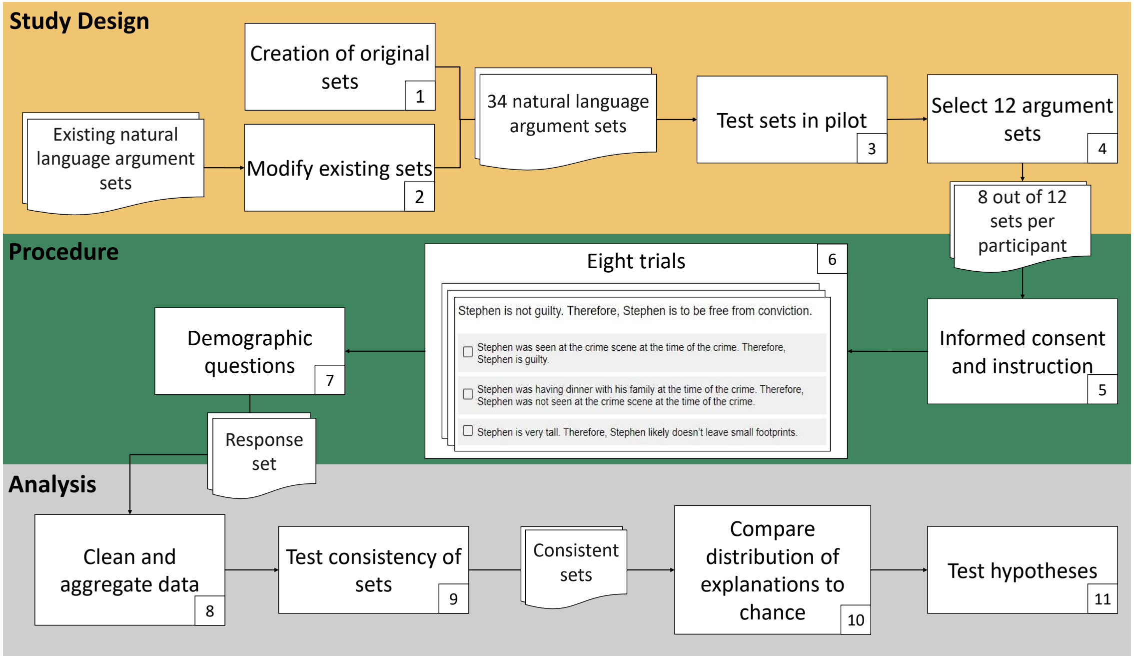


Figure 3: Overview of study design, procedure, and analysis.

4.1 Argument set creation and pilot study

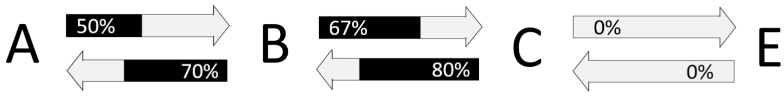
To test the two hypotheses, participants were tasked with selecting natural language arguments that explain some topic argument, where the natural language argument sets match the formal structure of the two AFs $\mathcal{AF}_1$ and $\mathcal{AF}_2$. For these natural language argument sets, we used arguments from prior work [1, 5, 6, 16] together with some newly added arguments (steps 1 and 2 in Figure 3). Each argument in the sets consists of one or two full sentences containing a premise and a conclusion. The conclusion of one argument can contradict the premise of another argument; this represents an undermining attack from the former to the latter (in line with [16]). Examples of these natural language argument sets are presented in Examples 3.1 and 3.2.

People may not interpret attacks in abstract argumentation instantiated using natural language as intended [5, 15]. Therefore, the natural language instantiations were validated using a pilot study (step 3 in Figure 3). In this pilot study, we compared the intended interpretation of argument sets (as $\mathcal{AF}_1$ and $\mathcal{AF}_2$) to the interpretations by participants (Figure 4). Five participants between the ages of 20 and 30 were presented with 34 argument sets. For each set, participants were presented with arguments in random order and asked to indicate the perceived attacks between arguments. It is important to note that the participants did not receive an explanation of formal argumentation or attacks.

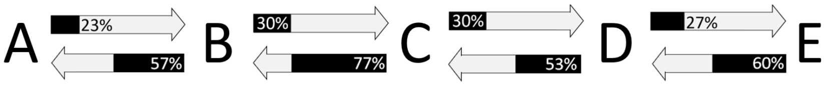
Participants generally saw unrelated arguments as unrelated to other arguments and related arguments as connected through attack relations. Participants frequently indicated bidirectional attacks, while attacks were in all cases intended to be unidirectional. Despite this, the intended attacks for both AFs were consistently among the most frequently selected attacks. This indicated that in most cases the participants agreed with the intended interpretation of the argument sets. The six argument sets that were most aligned with each AF based on the attacks indicated by participants were kept for the main study. For $\mathcal{AF}_1$ these were sets where argument E was unrelated. For $\mathcal{AF}_2$ these were sets where the entire path from E to A was found.[2] These 12 argument sets can be found in the online appendix[3] and the responses for these sets in Figure 4.

[2] In these sets, several attacks are perceived to be bidirectional. However, this is unlikely to cause issues in the experiment, since the topic argument (A) is presented as accepted and participants could not select the topic argument. Thus, they have to use the other arguments to select an explanation *given* that A is accepted.

[3] https://www.uu.nl/en/research/ai-labs/national-policelab-ai/related-explanations-in-formal-argumentation-an-empirical-study



(a) Attacks indicated for $\mathcal{AF}_1$



(b) Attacks indicated for $\mathcal{AF}_2$

Figure 4: Percentage that each attack between adjacent arguments was indicated in the pilot for the argument sets kept for the main study. Attacks between non-adjacent arguments indicated by participants are not shown in this figure for legibility, these attacks between non-adjacent arguments were all (except from E to C in $\mathcal{AF}_2$) indicated less than 10 per cent of the time.

4.2 Participants

For this study, convenience sampling used to recruit 127 participants from the argumentation mailing list, friends, family, and coworkers. The majority of participants (80) were between 18 and 35 years old, and most (104) completed at least a bachelor's degree. Participants rated their English reading proficiency at 4.2 out of 5 on average (SD = 0.75); therefore, we did not expect them to experience language-related difficulties. Prior familiarity with argumentation might influence study results; thus, participants were asked about their familiarity with formal argumentation. About half of the participants (63) reported not being familiar with or never having heard of formal argumentation. The other half were somewhat familiar (32) or very familiar to experts (33). Based on these descriptive statistics, participants were more familiar with argumentation than we expected the general population to be. We will return to this in the data analysis and discussion.

4.3 Procedure

Participants accessed the study through a digital link via the online survey tool Qualtrics.[4] At the start of the study, participants provided informed consent and received an explanation of the study procedure. The instructions given to participants can be found in the online appendix. Subsequently, the participants completed eight trials in random order, four corresponding to $\mathcal{AF}_1$ and $\mathcal{AF}_2$ each. Each trial featured one of the twelve natural language argument sets selected in the pilot. In such a trial, the topic argument (A) was presented at the top of the page. The other arguments in the set were presented below the topic in random order. An example of the layout is presented in the centre of Figure 3. The participants were instructed to explain why the conclusion of the topic arguments was the case using any combination of the other arguments. To construct their explanations, participants could select arguments by ticking a box next to them. They could include any number of arguments in their explanation, with a minimum of one. After the eight trials, participants were asked four demographic questions (see 4.2).

4.4 Analysis

After data collection, the data was analyzed in four steps, as outlined in the bottom row of Figure 3. Most of these steps involved preprocessing and grouping data so the hypotheses could be tested in the final step. In step 8, the data was cleaned and aggregated. This involved anonymising responses and removing incomplete responses. Additionally, potential issues related to response quality were assessed, such as giving the same answer to every question and responding extremely quickly. No such issues were found. The responses were then aggregated over all participants for each argument set.

In step 9, differences between the responses for the different natural language argument sets were investigated. There are two reasons such differences in responses could be found: some argument sets

[4] Version January 2023 of Qualtrics. Copyright 2024 Qualtrics, Provo, UT, USA. Available at https://www.qualtrics.com

were difficult for participants leading to more random responses, and some argument sets were, despite the pilot study, understood differently than intended. This would, for example, be the case if in one of the natural language argument sets for $\mathcal{AF}_2$ participants included E more frequently than in any other set. Differences of the first kind will not produce meaningful information and should be removed from the data, differences of the second kind, such as in the given example, could provide evidence against the hypotheses and should be included in the analysis.

In step 10, all responses were aggregated for the two hypotheses. The explanations per hypotheses were tested to determine whether variations in explanation count were statistically significant and not merely the result of chance. Such a significant difference between the actual explanation counts and the expected counts if participants answered randomly indicates a pattern in the data that warrants further analysis.

The fourth analysis step (step 11 in Figure 3) involved comparing participant explanations to the hypotheses. Based on the first hypothesis (*H1*), we expected participants to include only related arguments in their explanation. Based on $\mathcal{AF}_1$, this would mean that participants included C but not E in their explanations. For the second hypothesis (*H2*), we expected participants to prefer direct attackers over indirect attackers in the explanation of A in $\mathcal{AF}_2$. This would be confirmed if participants included C but not E in their explanations. Both hypotheses were tested by comparing the proportion of explanations that included C and excluded E to the proportion of explanations that included E.

5 Results

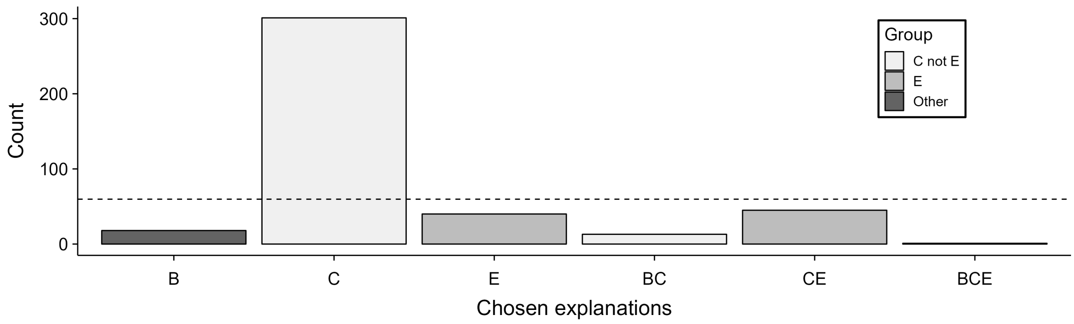
First, we investigated the differences between responses for the different natural language argument sets (step 9 in Figure 3). For the sets that instantiated $\mathcal{AF}_1$, we found the 4$^{\text{th}}$ and 6$^{\text{th}}$ set to deviate significantly ($\chi^2(6) = 30.34$, p <.001 and $\chi^2(6) = 45.82$, p <.001) from the total pattern of responses.[5] For set 6, $\{C, E\}$ was selected more than in other sets, and $\{C\}$ was selected less. Even though $\{C, E\}$ is not a related admissible explanation, it is an admissible extension, so this difference does not indicate any problematic interpretation issues. For set 4, all options but $\{C\}$ were selected more and $\{C\}$ was selected less. This indicates that participants responded more randomly, which was supported by feedback left on the study by participants, and thus this set was removed from further analysis. For $\mathcal{AF}_2$, set 4 and 6 also deviated significantly from the total pattern of responses ($\chi^2(6) = 61.71$, p <.001 and $\chi^2(6) = 45.81$, p <.001). Similarly to set 4 for $\mathcal{AF}_1$, these two sets showed a more random pattern of responses than the other sets and were thus removed.[6]

In the next step (10 in Figure 3), the remaining argument sets were grouped by hypothesis and the distributions of explanations were compared to chance (the dotted line in Figure 5). We found that explanations for both *H1* ($\chi^2(6) = 1168.20$, p <.001) and *H2* ($\chi^2(14) = 2253.90$, p <.001) differed significantly from the pattern that would be observed if participants answered randomly, indicating a pattern in responses that could be further analysed.

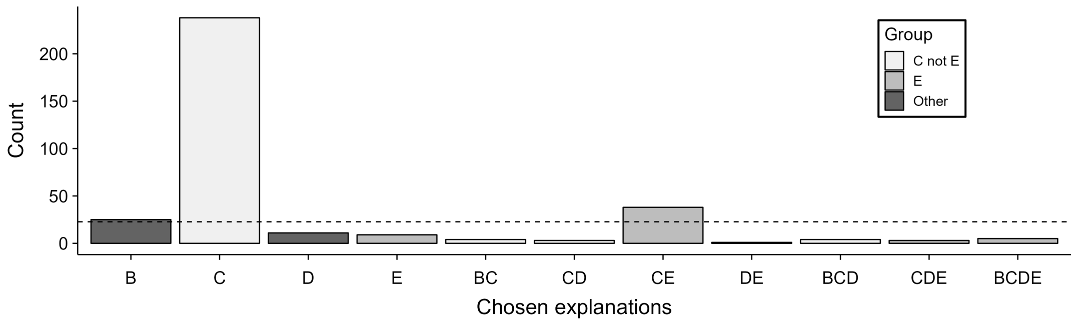
In the final analysis step (11 in Figure 3) the explanations were grouped to test the hypotheses. This resulted in three groups for each hypothesis, explanations supporting the hypothesis, which included C but not E, explanations including E, and other explanations. These groups are distinguishable by colour in Figure 5. The proportion of each of these groups of the total explanations was calculated, after which the proportions were compared using two-sample tests for equality of proportions. For *H1*, most of the provided explanations included C but not E and supported the hypothesis ($\hat{p} = .75$, $SE = 0.04$), a fifth of the explanations included E and undermined the hypothesis ($\hat{p} = .21$, $SE = 0.04$), and a small number of explanations fit neither group ($\hat{p} = .04$, $SE = 0.02$). Three two-sample tests for equality of proportions were conducted to test whether these proportions differed significantly from each other. More participants chose explanations aligning with the hypothesis than those not aligning with the hypothesis ($\chi^2(1) = 249.19$, p <.001) and explanations that fell into neither group ($\chi^2(1) = 437.74$, p <.001). The latter proportion was also lower than the proportion of explanations including E that undermined the hypothesis ($\chi^2(1) = 50.78$, p <.001). This, confirms the first hypothesis. Explanation $\{C\}$ was selected significantly more than all other

[5] All argument sets used and a full overview of statistical tests can be found in the online appendix, see footnote 3

[6] Two out of the three removed natural language argument sets are about persons and news articles being untrustworthy, which could indicate that people have a harder time interpreting the corresponding attack relations.



(a) Explanations selected for hypothesis *H1*



(b) Explanations selected for hypothesis *H2*

Figure 5: Total counts of the amount each explanation was selected for each of the hypotheses. Explanations that were never selected are omitted from the figure. The dotted line indicates the expected counts if participants answered randomly.

explanations,[7] so the high proportion of hypothesis-supporting explanations is driven by $\{C\}$ explanations.

For the second hypothesis, similar results were found. Most of the explanations provided included C but not E and supported the hypothesis ($\hat{p}$ = .71, SE = 0.04), a seventh of the explanations included E and undermined the hypothesis ($\hat{p}$ = .14, SE = 0.03), and a similar amount of explanations fit neither group ($\hat{p}$ = .15, SE = 0.03). Three two-sample tests for equality of proportions were conducted to test if these proportions were significantly different from each other. More participants chose explanations aligned with the hypothesis than those that undermined the hypothesis ($\chi^2(1) = 228.33$, p <.001) and than those explanations that fit neither group ($\chi^2(1) = 215.83$, p <.001). The proportion of explanations that fit neither group did not significantly differ from the proportion of explanations undermining the hypothesis ($\chi^2(1) = 0.30$, p = .587). This confirms the second hypothesis. Again, explanation $\{C\}$ was selected significantly more than all other explanations. Furthermore, the related admissible explanation $\{C,E\}$ was selected significantly more than all other explanations that included E.

5.1 Demographic effects

The effect of the collected demographic information on the given explanations was tested to support the generalizability of the results. No significant effects of age, education, or English proficiency were found. One small difference was found based on the participants' familiarity with argumentation. For $\mathcal{AF}_1$, self-identified experts selected $\{B,C\}$ significantly more than all other participants ($\chi^2(1) = 13.87$, p <.001). Experts might have considered the information in B necessary for C to serve as an explanation. No other

[7]The full comparisons of individual explanations can be found in the online appendix, see footnote 3 [8]

significant differences were found on the basis of familiarity with argumentation.

6 Discussion and outlook

This study aimed to empirically test the cognitive plausibility of two methods for selecting related arguments in argumentation-based explanations. This was done through an observational study in which participants provided explanations for the acceptance of topic arguments of two distinct AFs. We found that people primarily selected related and directly related arguments, respectively, when constructing explanations; this confirms both of our hypotheses and shows that relatedness in formal argumentation is cognitively plausible. The study employed a novel approach conducting the first empirical study into argumentation-based explanations.

The results found in the current study can be compared to those by [16], in which people tend to agree with reinstatement, that is a direct defender reinstates an attacked argument. In the current study, participants considered a direct defender the best explanation for the acceptance of an attacked argument. Additionally, people were able to identify this defender in a set of randomly ordered arguments, even without prior explanation of formal argumentation. This suggests that defence is a natural and intuitive concept.

Since this study is the first of its kind, we encountered several challenges in the methodology and generalizability, which we will highlight to guide future work in this field. One such challenge is that no other study has used unrelated arguments, which meant that we had to create these arguments. We created arguments that were not connected by attacks to the other arguments, but that did concern the same setting so as to not make the unrelatedness trivial (i.e. an argument such as "the moon is made of green cheese" is obviously unrelated to whether Dana can go ice skating) Despite pilot testing, some argument sets were excluded from the analysis because they were not properly understood by the participants.

One further challenge is that no compelling evidence was found that familiarity with argumentation affects explanations. However, in the sample, the participants had higher educational levels and were more familiar with argumentation than expected from the general population. Therefore, we should be careful when generalizing results to the entire population. An additional factor affecting the generalizability of this study is that we asked participants how they would choose to explain an argument, thus only conclusions about how people *construct* explanations can be drawn. Further research is needed to know what explanations people would prefer to *receive*. The findings of the current study are also limited to set-based acceptance explanations using abstract argumentation. This could be expanded in future work to include structured argumentation, different explanation methods, and non-acceptance explanations.

Another factor relevant to contextualizing this study is that, due to the design of the study, we do not know the motivation behind the choices of participants. To facilitate recruiting as many participants as possible we kept the experimental design simple, without any open questions about, for example, the reasoning behind decisions. Instructions to participants were also kept brief and did not include an explanation of argumentation or relatedness. Therefore, we cannot know why the participants chose the explanations they did and if any factors such as difficulty, fatigue, or confusion influenced their choices. Future research into the reasoning of participants to choose certain argumentation-based explanations over others would help to place the current study in a wider context and speculate on how results can be generalized to more complex argument structures and other argumentation formalisms.

In conclusion, this study found empirical evidence for argumentation-based explanation methods based on relatedness. These argumentation-based explanation methods [9, 3] are themselves inspired by research in social and cognitive science, formalizing how people select explanations. In this study, we have now validated these formalizations by showing that they align with how people select explanations. This can be expanded upon in future work by considering other argumentation-based explanation approaches, including structured argumentation and explanation for non-acceptance. And will hopefully inspire further investigations into argumentation-based explanations and possibilities of applied argumentation.

References

[1] E. Bezou Vrakatseli, H. Prakken, C. Janssen, L. Amgoud, and R. Booth. New experiments on reinstatement and gradual acceptability of arguments. In *Proceedings of the 19th International Workshop on Nonmonotonic Reasoning*, pages 109–118, 2021.

[2] G. A. Bodanza and E. Freidin. Confronting value-based argumentation frameworks with people's assessment of argument strength. *Argument & Computation*, 14(3):247–273, 2023.

[3] A. Borg and F. Bex. Minimality, necessity and sufficiency for argumentation and explanation. *International Journal of Approximate Reasoning*, 168:109143, May 2024.

[4] F. Cerutti, M. Cramer, M. Guillaume, E. Hadoux, A. Hunter, and S. Polberg. Empirical cognitive studies about formal argumentation. In D. Gobbay, M. Giacomin, G. Simari, and M. Thimm, editors, *Handbook of Formal Argumentation*, volume 2, page 851. College Publications, Aug. 2021.

[5] M. Cramer and M. Guillaume. Directionality of Attacks in Natural Language Argumentation. In C. Schon, editor, *Proceedings of the Fourth Workshop on Bridging the Gap between Human and Automated Reasoning*, volume 2261 of *CEUR Workshop Proceedings*, pages 40–46, Stockholm, Schweden, 2018. CEUR.

[6] M. Cramer and M. Guillaume. Empirical Study on Human Evaluation of Complex Argumentation Frameworks. In F. Calimeri, N. Leone, and M. Manna, editors, *Logics in Artificial Intelligence*, Lecture Notes in Computer Science, pages 102–115. Springer International Publishing, 2019.

[7] S. Doutre, T. Duchatelle, and M.-C. Lagasquie-Schiex. Visual Explanations for Defence in Abstract Argumentation. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 2346–2348, Richland, SC, May 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[8] P. Dung. 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.

[9] X. Fan and F. Toni. On computing explanations in argumentation. In B. Bonet and S. Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1496–1502. AAAI Press, 2015.

[10] M. Guillaume, M. Cramer, L. van der Torre, and C. Schiltz. Reasoning on conflicting information: An empirical study of Formal Argumentation. *PLOS ONE*, 17(8):e0273225, 2022.

[11] E. Hadoux and A. Hunter. Comfort or safety? Gathering and using the concerns of a participant for better persuasion. *Argument & Computation*, 10(2):113–147, 2019.

[12] B. Liao and L. van der Torre. Explanation semantics for abstract argumentation. In H. Prakken, S. Bistarelli, F. Santini, and C. Taticchi, editors, *Proceedings of the 8th International Conference on Computational Models of Argument (COMMA 2020)*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 271–282. IOS Press, 2020.

[13] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[14] S. Polberg and A. Hunter. Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *International Journal of Approximate Reasoning*, 93:487–543, 2018.

[15] H. Prakken and M. de Winter. Abstraction in argumentation: Necessary but dangerous. In S. Modgil, K. Budzynska, and J. Lawrence, editors, *Computational Models of Argument*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, pages 85–96. IOS Press, 2018.

[16] I. Rahwan, M. I. Madakkatel, J.-F. Bonnefon, R. N. Awan, and S. Abdallah. Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement. *Cognitive Science*, 34(8):1483–1502, 2010.

[17] A. Rosenfeld and S. Kraus. Providing Arguments in Discussions on the Basis of the Prediction of Human Argumentative Behavior. *ACM Transactions on Interactive Intelligent Systems*, 6(4):1–33, 2016.