Relevant Explanations in Formal Argumentation, an Empirical Study

Rosalie Johanna Scheffers
r.j.scheffers@student.uu.nl
9559035

Project supervisor: Prof. dr. F.J. Bex
Second examiner: Dr. A. Borg

A thesis presented for the Artificial Intelligence master

Utrecht University

Graduate School of Natural Sciences (GSNS)
Utrecht University
The Netherlands
May 26, 2023

Abstract

The use of automated decision-making is becoming increasingly prevalent. Users of systems that make these decisions must be able to assess a system's biases and have trust in it. Providing explanations for system decisions is one way to achieve this. Providing these explanations is the focus of the Explainable Artificial Intelligence (XAI) field. One technique used within XAI is formal argumentation. The logic used by an algorithm to arrive at a specific decision can be represented via formal argumentation structures. However, how such an argumentation structure can be translated into human-friendly explanations remains an open question. One concept formalized for explanations in argumentation that takes into account properties of human explanations is 'relevance'. Informally an argument is relevant to another argument if there is a relation between the two, for example, by attacking or defending an argument.

In this thesis, the concept of relevance was empirically tested by comparing explanations in formal argumentation based on relevance to explanations provided by participants. One hundred twenty-seven participants provided explanations for scenarios based on two different types of relevance. Based on the results, relevance in argumentation seems to align with explanations selected by participants. Participants preferred small explanations consisting of direct defenders, arguments that attack the attacker of an argument. However, further investigation is needed to determine whether the task's difficulty affects this study's results. Future work could build on the current work by expanding to non-acceptance and non-extension-based explanations and by investigating differences in explanation behaviour based on prior knowledge and goals of explanation.

Contents

1 Introduction

Explainable artificial intelligence (XAI) is a field that seeks to develop artificial intelligence (AI) models that are transparent and interpretable to humans. In the context of AI, a model is a mathematical or computational representation of a system or process used to make predictions based on input data. Understanding how these models make decisions is increasingly crucial as AI is integrated more deeply into society. AI is being used in a wide range of applications, including healthcare, finance, and criminal justice, where the decisions made by these systems can have a significant impact on individuals and society. The purpose of XAI is to provide a means for human users to comprehend the decision-making process used by AI systems. This includes understanding the inputs and outputs of the system, as well as the underlying algorithms and models. XAI is vital because it can help increase AI systems' transparency, accountability, and trustworthiness, which are critical factors in their adoption and use (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Tintarev & Masthoff, 2007). A variety of approaches are being used in the field of XAI. These range from more straightforward methods, such as model visualization and feature importance analysis, to more complex techniques, such as formal argumentation and counterfactual analysis.

To provide people with the best possible explanation of an AI system or decision, it is important to consider that explanation is an argumentative and social process (Miller, 2019; Mothilal, Sharma, & Tan, 2020). People attribute human-like qualities to AI systems; considering these qualities can improve explanations (Miller, 2019). One method that can take into account social aspects of explanation is formal argumentation, which represents model decision-making using a formal graph structure comprised of arguments and attacks between arguments.

This graph structure is called an argumentation framework (AF) and can capture the argumentative nature of a human conversation in which parties provide arguments and counterarguments. Argumentation frameworks are used in XAI as post-hoc or intrinsic models. Post-hoc models are argumentation frameworks derived from non-argumentative models. Intrinsic models are natively created and used as argumentative models (Čyras, Rago, Albini, Baroni, & Toni, 2021). Both of these types of models give insight into the decision-making process. To provide explanations for individual decisions or conclusions, sets of arguments can be selected from these models. The most basic way to select such arguments is based on the argumentation semantics introduced by Dung (1995). These semantics identify sets of arguments which together are 'acceptable'. This basic method for selecting explanations takes into account the argumentative nature of explanations. However, there are several other important qualities of explanation that can and should be taken into account, such as causality, selectiveness, and counterfactuality (Miller, 2019). Several methods have been proposed for selecting explanations that take these properties into account (Borg & Bex, 2021a, 2021b; Fan & Toni, 2014). These methods formalize selectiveness and causality as different criteria for including or excluding arguments in explanations. One concept formalized for explanations in argumentation that supposedly aligns with human cognitive biases is 'relevance' (Borg & Bex, 2021b). Informally an argument is relevant to another argument if there is a link between the two in an AF, for example, by attacking or defending the argument (Fan & Toni, 2014).

The proposed methods for selecting explanations formalize qualities of human explanations; however, no method of explanation based on formal argumentation has been evaluated using human participants. It is essential to compare explanations in explainable artificial intelligence to explanations given by people. One metric for comparing something to human behaviour is cognitive plausibility. Explanations given by an AI system are cognitively plausible if they align with human explanations(Kennedy, 2009). This work is the first to test the cognitive plausibility of explanations based on formal argumentation.

In this thesis, this concept of 'relevance' will be empirically tested by comparing explanations in formal argumentation based on this concept to real explanations provided by people to answer the following research question.

Are explanations based on relevance in formal argumentation cognitively plausible?

This question will be answered using a non-experimental observational study. In this study, participants were presented with sets of arguments, each of which was a natural language instantiation of an argumentation framework. Each set included a topic and several other arguments. Participants were tasked to choose from the other arguments in the set to explain the topic argument. After the study, explanations chosen by participants were compared to explanations based on formal argumentation theory to determine whether answers given by participants aligned more closely with explanations that were or

were not based on relevance.

The methods employed in this study build on previous research into the relationships between various components of formal argumentation and human cognition. The initial investigation in this domain was conducted by Rahwan, Madakkatel, Bonnefon, Awan, and Abdallah (2010), who examined and found evidence for the cognitive plausibility of reinstatement. Subsequently, other studies have delved into rebuttal (Yu, Xu, & Liao, 2018), attack relations (Cramer & Guillaume, 2018a), and semantics (Guillaume, Cramer, van der Torre, & Schiltz, 2022). In this study, the methods and arguments sets were influenced by and based on the works of Guillaume et al. (2022) and Rahwan et al. (2010).

This study's contributions to empirical research in computational argumentation are twofold. First, this study attempts to bridge the gap between theoretical and practical applications of formal argumentation for XAI. The research question's resolution will shed light on how formal argument can be used within XAI by offering an empirical foundation for selecting between methods for generating explanations. Suppose the study discovers that participants' explanations align with explanations in formal argumentation based on relevance. In that case, this may indicate that relevance is a crucial factor when generating explanations using formal argumentation. This could provide guidance for the creation of more effective approaches for generating XAI explanations that are more closely aligned with human reasoning.

A second contribution of this study is the exploration and use of a novel method for investigating explanations based on formal argumentation. The novelty of this method lies in the observational study design in which participants select their explanations. In previous studies (Guillaume et al., 2022; Rahwan et al., 2010), participants evaluated existing arguments but did not have the freedom of choice present in the current study, which allows for a more natural observation of human behaviour. The development of this method facilitates further research into the connections between formal argumentation and real-world explanations. This could ensure that XAI systems are developed and evaluated in a more rigorous and scientifically sound manner by comparing them directly to human explanations, potentially enhancing their reliability and effectiveness in real-world applications.

2 Explanations and Argumentation

This section will present the theory relevant to the research question. First, a short overview of XAI will be presented, which will include some important terms and definitions and motivation for doing research in this field. This will be followed by a section on explanations, including the different goals and desirable features explanations can have. Finally, different methods for forming explanations based on formal argumentation will be discussed.

2.1 Explainable artificial intelligence

Explainable artificial intelligence (XAI) is a subfield of AI research concerned with the interpretability of AI systems. The degree to which a person understands the cause of a decision is referred to as interpretability (Miller, 2019). This can be accomplished by designing interpretable systems or by making systems more interpretable through explanations. Early Artificial Intelligence was interpretable because it was traceable. All decision-making steps could be followed from the input to the output because systems were only of limited complexity (Longo, Goebel, Lecue, Kieseberg, & Holzinger, 2020). As more data has become available, machine learning and statistical methods have become more popular (Angelov, Soares, Jiang, Arnold, & Atkinson, 2021). As AI and machine learning have become more widespread and complex, models have become harder to trace. In these models, it is more difficult to interpret how information is used to make decisions. Therefore, there have been increasing concerns about bias, fairness, and representation. This has led to an increase in research towards model interpretability (Angelov & Soares, 2020; Confalonieri, Coba, Wagner, & Besold, 2021). The goal of XAI is to construct AI systems that are both highly accurate and easy to interpret (Angelov et al., 2021). This can be done through explanations that bridge the gap between AI systems and those who use them by providing insight into the decision-making process of a system (Čyras et al., 2021).

In XAI, automated decisions can be explained for various reasons. These reasons can be divided into three categories based on the target audience of the explanation. These are explanations for laypeople, explanations for professional users, and explanations for system developers. Laypeople who interact with a system once or a handful of times prefer explanations for specific events and decisions (Miller, 2019). Explanations for laypeople are often aimed at improving understanding of decisions, finding meaning in decisions, and persuasion of the correctness of decisions (Miller, 2019). Another important reason to explain system decisions to laymen is the 'Right to Explanation', which is, according to some, included in the European Union GDPR (Gacutan & Selvadurai, 2020; Selbst & Powles, 2017; Winikoff & Sardelić, 2021). According to articles 13 through 15, individuals have the right to "meaningful information about the logic involved" in automated decisions which have legal or otherwise significant effects on them. According to Selbst and Powles (2017), this constitutes a right to explanation. Professional users are more familiar with a system than laypeople and have different reasons for desiring explanations. Professional users interact more with the system and want to know how the system functions to get the most out of their interactions. Professionals should require less explanation as they interact more with a system (Miller, 2019). Having insight into the system through explanations allows professional users to aid in improving the system, creating appropriate regulations, and interpreting and adding context to system decisions. Explaining decisions can increase a user's trust in a system (Dzindolet et al., 2003; Tintarev & Masthoff, 2007). Mistakes made by a system reduce users' trust in the system unless an explanation is provided (Dzindolet et al., 2003). The third category, developers of a system, want explanations to gain insight into how to improve a system and to verify that the system is working as intended (Confalonieri et al., 2021; Verma, Lingenfelder, & Klakow, 2020). It is important to be able to verify the working of a system because developers are responsible for decreasing bias and discrimination and increasing the fairness of a system. This is especially important for systems used in sensitive domains, such as healthcare or police work, to ensure the safety of systems (Verma et al., 2020). Increased interpretability and a better understanding of these models will allow them to be more widely used (Longo et al., 2020).

There are three main approaches to explanations in XAI. *Global methods* concern themselves with explaining the overall functioning of a model (Confalonieri et al., 2021). These methods do not explain individual decisions but can be used to see large-scale differences between groups of input. *Local methods* explain specific decisions made by a model. Explanations focus on a specific case, and explanations of one

model can vary greatly depending on the case (Ribeiro, Singh, & Guestrin, 2016). *Surrogate models* are models trained to explain the decisions of other models. This can be done with local or global methods (Čyras et al., 2021). When a model is not interpretable to an audience, a surrogate model can provide understandable explanations to bridge the gap between people and an AI system (Angelov et al., 2021; Gilpin et al., 2018). Opaque models always need this extra explanation step to be turned into explainable models.

2.2 Explanations

As discussed in the previous section, being able to provide explanations of system decisions is a desirable property of systems. However, there is no clear consensus on what such an explanation should look like or what would make for a good explanation. In psychology, explanation involves assigning causal attribution to events (Longo et al., 2020; Miller, 2019). In philosophy, reasoning and explanation are argumentative activities (Mercier & Sperber, 2011). And in computer science, surrogate models are evaluated using fidelity, the extent to which they stay true to the model being explained (Confalonieri et al., 2021). To automatically generate explanations and to be able to evaluate these explanations, it is vital to understand how people define, create, and evaluate explanations by integrating information from various fields (Verma et al., 2020). These fields have very different but not necessarily incommensurable views on explanation. In their 2020 study, Verma et al. propose a definition of explanation in an AI context that aims to integrate elements from different disciplines:

> *An explanation is a representation of fair and accurate assessments made by an explainer to transfer relevant knowledge (about a given scenario) from the explainer to a recipient* (Verma et al., 2020).

The individual components of this definition require some further definition. *Representation* can be done using various media such as text, dialogue, images, or a combination. This representation must be an appropriate assessment choice by the explainer and acceptable to the recipient. *Assessments* refer to either observations or the analysis of observations. In an AI context, this will generally refer to the inner working of a system and the factors of the input most relevant to the output. *Fair and accurate* refers to observations that are unbiased and have high fidelity to the system. The *explainer* is the entity providing the explanation; in AI, this is generally a model or system which selects an explanation out of all possible causes and communicates this to the recipient (Miller, 2019). The *recipient* is presented with a clarification on the output of a system or in response to a posed question. Finally, the *transfer of relevant knowledge* entails information gained either actively or passively by the recipient, which is necessary to accomplish the goals of the explainer and the recipient. The following section will discuss the goals of explanations in more detail.

2.2.1 Goals of explanations

Both the explainer and the recipient can have goals for an explanation. It is important to consider different goals of explanations since they may be evaluated based on different criteria. According to Antaki and Leudar (1992), an explanation is given based on a posed question by the recipient. This question can be implicitly or explicitly stated and indicates the recipient's goal by indicating what they want to know. Miller (2019) divides these explanatory questions into three categories. *What-questions*, such as "What event happened?", *how-questions*, such as "How did that event happen?", and *why-questions*, such as "Why did that event happen?".

The goals of the explainer can depend on the question posed by the recipient, the limitations of the explainer, and further motivations of the explainer, such as convincing the recipient to continue using the system (Antaki & Leudar, 1992). Tintarev and Masthoff (2007) define seven different goals for explanations in recommender systems. These can be found in Table 1. If the explainer is interested in getting the recipient to buy an item, the explanation might have as a goal to be persuasive. To have the recipient return to use the system in the future, it can be desirable to deliver a trustworthy explanation (Confalonieri et al., 2021). Efficient and effective explanations can help users make fast and sound decisions making a system easier to use (Tintarev & Masthoff, 2015).

The goals of explanations are not independent of each other. In an experiment by Balog and Radlinski (2020), participants were asked to generate explanations based on each of the seven goals from Tintarev and Masthoff (2015). Then a different group of participants were asked to rate these explanations based

Goal	Definition	How to measure
Transparency	Explain how the system works	Ask users, measure user time and accuracy
Scrutability	Allow users to tell the system it is wrong	Users can identify mistakes in the system
Trust	Increase users' confidence in the system	Loyalty and increased usage
Effectiveness	Help users make good decisions	Small change in rating of items
Persuasiveness	Convince users to try or buy	Difference in the likelihood of selecting item
Efficiency	Help users make decisions faster	Completion time of using system
Satisfaction	Increase the ease of usability or enjoyment	Ask users, measure loyalty

Table 1: Seven different aims of explanations as defined by Tintarev and Masthoff (2015), suggestions for how to measure have been added based on Confalonieri et al. (2021) and Balog and Radlinski (2020).

on the goals. They found that when people generate an explanation focusing on a specific goal, this does not mean this will be the aim others consider most prevalent in the explanation. Transparency aids trust and effectiveness; aiming for trust enhances scrutability and transparency. Based on these relations between goals Balog and Radlinski (2020) suggest that satisfaction, scrutability, and transparency may provide the most complete assessment of the quality of an explanation across the seven goals.

2.2.2 Desirable features of explanations

A well-known human cognitive bias is anthropomorphizing non-human entities by assigning them beliefs, desires, and intentions. Such biases can influence how people interact with XAI systems because they are more likely to attribute human-like qualities to these systems and expect them to behave similarly to humans. To meet the expectations of the recipient of an explanation, it is necessary to consider this bias when developing XAI systems (Miller, 2019). To meet these expectations, explanations should have the following characteristics.

- *Causal.* When asking for an explanation, people frequently are interested in the causes for an event or decision (Miller, 2019). AI systems that seem to make decisions based on causal relations should also be able to provide explanations that refer to these causal elements (Longo et al., 2020).
- *Counterfactual.* People seek explanations in response to counterfactual or contrastive cases, especially when encountering similar events with different outcomes. People want to know the difference between them (Miller, 2019; Mothilal et al., 2020). Mothilal et al. (2020) argue that the explainer should choose an *actionable* counterfactual explanation. This means the recipient should be able to do something with the explanation.
- *Argumentative.* People reason to create and evaluate arguments meant to persuade others and defend their beliefs and opinions (Mercier & Sperber, 2011). Therefore, normal conversations are argumentative. Thus explanations provided by a system should be, too, since that is what people expect (Čyras et al., 2021)
- *Social.* Explanations should consider the background, desires, and intentions of the recipient. The explainer should also ensure the recipient understands the explanation, which can involve tailoring the explanation to the recipient (Miller, 2019).
- *Selective.* When people give explanations, they tend only to mention a small subset of all possible causes, which are selected in a biased manner. One such bias is that people tend to focus on causes consistent with prior beliefs (Miller, 2019). Explanations should be selected to be relevant to the recipient and not include the entire cause of an event.
- *Interactive.* Recipients should be able to interact with an explanation to get relevant information. There are many properties of recipients that influence what explanation they would prefer. According to Confalonieri et al. (2021), lay users may be more interested in understandability than accuracy, while expert users might prefer more precise explanations. Since not all recipients are interested in the same explanation, it can help to provide an interactive explanation or multiple explanations. This can also increase the recipient's trust in the explanation (Arrieta et al., 2020; Balog & Radlinski, 2020).

2.3 Argumentation for explanation

Formal argumentation theory is a method for non-monotonic reasoning based on constructing and evaluating arguments within an argumentation framework (AF). The structure of AFs allows the representation of information and decision-making in a way that facilitates the generation of explanations from the structure. AFs are able to handle conflicting and defeasible information. This allows AFs to model the argumentative and fallible character of human reasoning (Atkinson et al., 2017). AFs also model several of the important characteristics for explanation discussed in the previous section. AFs can facilitate argumentative and interactive dialogue, provide a basis for selecting the most important causes, and provide contrastive cased (Vassiliades, Bassiliades, & Patkos, 2021; Miller, 2019). This makes AFs interesting for XAI. Čyras et al. (2021) distinguish two approaches in which AFs are used for explanation in XAI. *Intrinsic* approaches, where models natively use argumentative techniques, and *post-hoc* methods, where an argumentation framework is derived from a non-argumentative model.

Post-hoc models can either be *complete* or *approximate* representations of the explained model. In the complete post-hoc approach, the AF represents the entire non-argumentative model. This approach has been used for scheduling by Čyras, Letsios, Misener, and Toni (2019). The authors translated the decisions of users in scheduling problems into an AF to extract explanations for why a schedule is infeasible or inefficient. In the approximate post-hoc approach, explanations rely upon incomplete mappings between the non-argumentative model and an AF. This is used in cases where it is not possible or useful to capture the entire non-argumentative model using an argumentative model. One application of the approximate post-hoc approach to Bayesian Networks (Timmer, Meyer, Prakken, Renooij, & Verheij, 2017) uses structured argumentation frameworks to model and explain the interplay between variables by focusing on specific variables and interactions and not the entire model.

Intrinsic AF approaches are popular in recommender systems, where the AF can be used both to generate and explain a recommendation. One such approach, (Briguez et al., 2014), uses DeLP to handle incomplete and contradictory information for movie recommendations. Oren, van Deemter, and Vasconcelos (2020) uses ASPIC$^+$ to create an explainable argument-based system for planning. Furthermore, Odekerken, Bex, Villata, Harašta, and Křemen (2020) created a human-in-the-loop intrinsic argumentation system which is currently being implemented for the classification of fraudulent webshops at the Dutch Police.

Post-hoc models serve as interpretable representations of non-argumentative models, while intrinsic models are intentionally designed with interpretability in mind. However, these models alone do not provide explanations for individual decisions. To explain specific model decisions, a subset of arguments within the model must be selected. This work focuses on providing explanations for arguments in AFs, offering context for accepting or rejecting an argument. Such explanations focus on a single decision or outcome of a system. In this study, simple intrinsic models are used to select arguments for explanations, although this approach could also be applied to post-hoc models. Before specific strategies for generating explanations can be discussed, an introduction to formal argumentation is required.

Formal argumentation encompasses both abstract and structured argumentation approaches. Abstract argumentation involves representing arguments as abstract entities and defining a set of rules for reasoning about their relationships. The focus is on identifying the most plausible set of arguments rather than on their specific structure. Structured argumentation, on the other hand, involves representing arguments as structured objects and defining a set of rules for manipulating and evaluating them. The focus is on the structure and content of arguments rather than on their abstract properties. The following sections will discuss the most important terms and definitions in both abstract and structured argumentation. Then these terms and definitions will be used to discuss how formal argumentation can be used to generate explanations.

2.3.1 Abstract argumentation

In abstract argumentation, arguments have no internal structure. The premises and conclusions of arguments are not represented. The nature of the attack relation in abstract argumentation is unspecified; there is no difference between an argument which is attacked on a premise and an argument attacked on its conclusion. These abstract notions of arguments and attacks are represented in argumentation frameworks. *Abstract argumentation frameworks* (Dung, 1995) are pairs $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments, and $\mathcal{R}$ is a binary attack relation on these arguments. For each tuple $(A, B) \in \mathcal{R}$, A is the attacking argument. An example of an AF can be seen in Example 2.1. An AF can be represented as a

directed graph, in which arguments are represented by nodes and attack relations by arrows between the nodes, an example of this can be seen in Figure 1.

In this section, the basic properties of AFs will be discussed, but it is worth noting that there are several approaches which expand on Dung-style AFs. *Bipolar argumentation frameworks* (BAFs) include a *support* relation between arguments ($\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$) (Cayrol & Lagasquie-Schiex, 2005). In *probabilistic argumentation frameworks*, Dung-style AFs are extended with probability functions on (set of) arguments (Hunter & Polberg, 2017). And in *preference-based argumentation frameworks* (PAFs), preference relations between arguments are added, and arguments cannot be defeated by attackers with a lower preference (Amgoud & Cayrol, 1998).

Example 2.1 Figure 1 represents $\mathcal{AF}_1 = (\mathcal{A}_1, \mathcal{R}_1)$, where $\mathcal{A}_1 = \{A, B, C, D, E\}$ and $\mathcal{R}_1 = \{(A, B), (C, B), (C, D), (D, C), (D, E), (E, E)\}$.

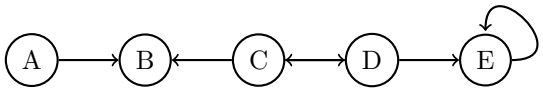


Figure 1: A graphical representation of $\mathcal{AF}_1$.

Given an AF, arguments that are *acceptable* can be determined by applying Dung-style semantics (Dung, 1995). This can be done using argument extensions. In this extension-based approach, sets of arguments (or *extensions*) that are collectively acceptable are identified.

Definition 2.1 (Acceptability) An argument A is *acceptable* with respect to a set of arguments S iff each argument attacking A is attacked by S. When A is acceptable with respect to S, S defends A.

Using this definition of acceptability, several semantics have been introduced by Dung (1995):

- An *admissible extension* is conflict-free, i.e. no argument in the extension attacks an argument in the extension, and each argument in the extension is acceptable with respect to it;
- A *complete extension* is an admissible extension that contains all arguments it defends;
- The *grounded extension* is the minimal (w.r.t $\subseteq$) complete extension;
- A *preferred extension* is a maximal (w.r.t $\subseteq$) complete extension;
- A *stable extension* is a complete extension that attacks all arguments not in the extension.

In $\mathcal{AF}_1$ from Example 2.1 the grounded extension is $\{A\}$, the preferred extensions are $\{A, C\}$ and $\{A, D\}$, and the stable extension is $\{A, D\}$.

Multiple other semantics have subsequently been introduced, such as semi-stable (Verheij, 1996; Caminada, 2006), ideal (Dung, Mancarella, & Toni, 2007), and stage semantics (Verheij, 1996).

Definition 2.2 Let $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$ be an argumentation framework. Then for $S \subseteq \mathcal{A}$, S^+ is the set of all arguments attacked by S.

- A *semi-stable extension* is a complete extension such that $S \cup S^+$ is maximal (w.r.t $\subseteq$);
- The *ideal extension* is the maximal (w.r.t $\subseteq$) admissible set that is a subset of each preferred extension;
- A *stage extension* is conflict-free and $S \cup S^+$ is maximal (w.r.t $\subseteq$) among conflict-free sets.

One additional semantics to consider is CF2 semantics, first introduced by Baroni and Giacomin (2003). CF2 semantics is interesting because it seems to align well with human reasoning (Cramer & Guillaume, 2019; Guillaume et al., 2022). This paper will introduce some basic intuitions behind CF2 semantics; for the full definition, see Baroni and Giacomin (2003) or Gaggl and Woltran (2013) for a revised version. The idea behind the semantics is that an AF is partitioned into strongly connected components (SCC). Each component is recursively evaluated by choosing maximal conflict-free sets in each component and removing arguments attacked by chosen arguments (Cramer & van der Torre, 2019). A subgraph of an AF is an SCC if there is a path from each argument to every other argument

in the subgraph. More detailed definitions and an ontology of argumentation semantics can be found in (Baroni, Caminada, & Massimiliano, 2018).

Besides semantics, the acceptance of an argument is also determined by *acceptance strategies*. When faced with a conflict between arguments, one can take a sceptical approach and not accept any arguments involved in the conflict. Alternatively, one can take a credulous approach and accept an argument as soon as there is evidence for it. Formally these concepts can be defined as follows:

Definition 2.3 (Credulous and sceptical acceptance) Let $AF = (\mathcal{A}, \mathcal{R})$ be an argumentation framework and `Sem` a semantics, then $A \in \mathcal{A}$ is

- *Sceptically accepted* ($\cap$) iff A is part of all extensions of `Sem`; and
- *Credulously accepted* ($\cup$) iff A is part of at least one extension of `Sem`.

Abstract argumentation frameworks are valuable models to study fundamental argumentation mechanisms, prove results for large classes of systems, and investigate the commonalities and differences between existing non-monotonic logics (Dung, 1995; Prakken & de Winter, 2018). However, there are some dangers to only focusing on abstract argumentation without considering the structure and information in arguments. The first concern expressed both by Caminada and Wu (2011) and Prakken and de Winter (2018) is that modelling natural language examples of argumentation frameworks directly in abstract argumentation frameworks ignores the important step of consulting a theory of the nature of arguments and attacks. Another concern raised by Prakken and de Winter (2018) is that AFs are often made within a specific context but are, after abstraction, generalised to other contexts. Thus it is important when instantiating AFs to have a clear account of the nature of arguments and attack relations and the context in which argumentation occurs (Prakken, 2010). One approach to this is to always start with a fully specified system before abstracting ways from it instead of the other way around (Caminada & Wu, 2011). To specify systems in more detail, by considering the nature of attacks and the internal structure of arguments, an account of structured argumentation will be given in the next section.

2.3.2 Structured argumentation

In structured argumentation, a formal language for representing knowledge is assumed, and that knowledge is used to specify arguments. These arguments are structured because they have a premise, conclusion, and formal relation between them. The attack relation is based on these individual elements of arguments. Three approaches to structured argumentation are ABA (Bondarenko, Dung, Kowalski, & Toni, 1997), ASPIC$^+$ (Prakken, 2010), and Defeasible Logic Programming (DeLP) (García, Chesñevar, Rotstein, & Simari, 2013). In ABA, each argument corresponds to a set of assumptions that proves a claim. An argument attacks another argument if its claim contradicts an assumption of the other argument. In both ASPIC$^+$ and DeLP, arguments are constructed using strict and defeasible rules.

ASPIC$^+$ (Prakken, 2010), an instantiation of Dung-style AFs discussed in the previous section, will be discussed in some more detail. ASPIC$^+$ abstracts away from the nature of logical languages and can be instantiated in various ways. ASPIC$^+$ can also incorporate argument orderings based on preferences (Besnard et al., 2014; Prakken, 2018).

An implementation of ASPIC$^+$ requires an *argumenation system* (AS) and a *knowledge base* ($\mathcal{K}$), which together form an *argumentation theory* (AT). The

Definition 2.4 (Argumentation system) An *argumentation system* is a tuple $AS = (\mathcal{L}, \mathcal{R}, n)$ where:

- $\mathcal{L}$ is a logical language closed under negation ($\neg$), we denote $\psi = -\varphi$ if $\psi = \neg\varphi$ or $\varphi = \neg\psi$.
- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible rules ($\mathcal{R}_d$) of the form $\{\varphi_1, ..., \varphi_n\} \rightarrow \varphi$ or $\{\varphi_1, ..., \varphi_n\} \Rightarrow \varphi$ respectively, such that $\{\varphi_1, ..., \varphi_n, \varphi\} \subseteq \mathcal{L}$ and $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$. Here $\{\varphi_1, ..., \varphi_n\}$ are the *antecedents* and φ the *consequent* of the rule;
- n is a naming function such that $n : \mathcal{R}_d \rightarrow \mathcal{L}$, such that $n(r)$ is a well-formed formula in $\mathcal{L}$ which says that $r \in \mathcal{R}_d$ is applicable.

Argumentation systems use strict and defeasible rules; informally, in strict rules, the consequent follows *without exception* from the antecedent, and in defeasible rules, the consequent *presumably* follows from the antecedent.

Definition 2.5 (Knowledge base) A *knowledge base* in an argumentation system is a set of formulas $\mathcal{K} \subseteq \mathcal{L}$ which contains two disjoint subsets:

- $\mathcal{K}_n$, the set of *axioms* which cannot be questioned and;
- $\mathcal{K}_p$, the set of *ordinary premises* which can be questioned.

In ASPIC$^+$, arguments are defined relative to an argumentation theory. And can be constructed step-by-step by chaining inference rules starting with elements in the knowledge base.

Definition 2.6 (Argument) An argument A, on the basis of an argumentation theory AT with a knowledge base $\mathcal{K}$ and an argumentation system AS, is:

- φ if $\varphi \in \mathcal{K}$, or
- $A_1, ..., A_n \rightarrow \psi$, if $A_1, ..., A_n$ are arguments such that there exists a strict rule $\texttt{Conc}(A_1), ..., \texttt{Conc}(A_n) \rightarrow \psi$ in $\mathcal{R}_s$, or $A_1, ..., A_n \Rightarrow \psi$, if $A_1, ..., A_n$ are arguments such that there exists a defeasible rule $\texttt{Conc}(A_1), ..., \texttt{Conc}(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$.

Arguments contain sub-arguments which support intermediate conclusions. Components and properties of arguments can be retrieved using functions. The function `Prem` returns all *premises* used to build an argument, `Conc` returns the conclusion of an argument, `Sub` returns all sub-arguments, `DefRules` returns all defeasible rules used in the construction of an argument, and `TopRule` returns the last rule used in the argument.

Attacks on arguments are based on the rules and premises applied in the construction of an argument. If these rules are defeasible or premises are ordinary, they can be attacked with the following attack types.

Definition 2.7 (Attack) *A attacks B* iff *A undercuts*, *rebuts*, or *undermines B*, where:

- *A undercuts B* (on B') iff $\texttt{Conc}(A) = -n(r)$ for some $B' \in \texttt{Sub}(B)$ such that B''s top rule r is defeasible;
- *A rebuts B* (on B') iff $\texttt{Conc}(A) = -\varphi$ for some $B' \in \texttt{Sub}(B)$ of the form $B''_1, ...B''_n \Rightarrow \varphi$;
- *A undermines B* (on ψ) iff $\texttt{Conc}(A) = -\varphi$ for an ordinary premise φ of B.

2.3.3 Argument explanation

In this work, explanations of arguments in AFs entail providing context for why an argument is or is not accepted. A basic method to provide context for the acceptance of an argument based on a semantics is to return one or all extensions the argument is part of. However, this basic method does not necessarily provide satisfying explanations for people since these explanations are very general and do not consider the properties discussed in Section 2.2.2, such as selectiveness, causality, and counterfactuality. Several researchers have proposed methods for explaining arguments that attempt to take into account desirable properties of explanations (Miller, 2019), such as causality in related admissible explanations (Fan & Toni, 2014) and selectiveness in verbose and compact explanations (Fan & Toni, 2015a). Other common forms of explanations based on AFs use dialogue games or sub-graphs; these will not be considered in this work because empirical work on testing explanations for humans has focussed on extension-based methods. Comprehensive overviews of argumentation methods for explanation in XAI can be found in Čyras et al. (2021) and Vassiliades et al. (2021). In the following section, only those methods for argument explanation that relate to desirable properties of explanations as discussed in Section 2.2.2 will be discussed.

Related admissibility is a semantics for abstract argumentation proposed by Fan and Toni (2014). This semantics is based on the idea that arguments that do not contribute to the acceptance of the topic should be excluded from the explanation for that topic. This notion of related admissibility shares similarities with the properties of causality and selectiveness in explanations (Miller, 2019). In a related admissible extension, all arguments are connected to the topic, and they can be seen as the reasons for accepting the topic argument of the explanation. In other words, the arguments that are part of a related admissible extension can be seen as arguments selected because they are causally relevant to the acceptance of the topic argument.

To further define related admissible semantics, a *defends* relation is used.

Definition 2.8 (Defends) (Fan & Toni, 2014, Definition 1.) An argument B *defends* an argument A iff:

1. B is A; or
2. B attacks C and C attacks A; or
3. B defends C and C defends A.

Then using this defends relation *related admissibility* is defined.

Definition 2.9 (Related admissible) (Fan & Toni, 2014, Definition 2.) Let $\mathcal{AF} = (\mathcal{S}, \mathcal{R})$ and $S \subseteq \mathcal{A}$. S is *related admissible* iff there is an argument $A \in S$ such that A is defended by every argument in S and S is admissible. Argument A is the *topic* of S.

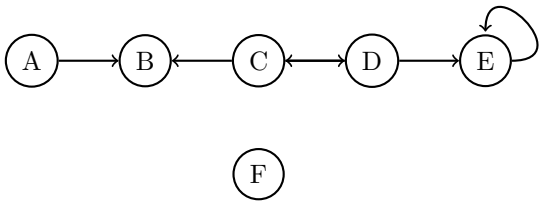


Figure 2: A graphical representation of $\mathcal{AF}_2$.

In $\mathcal{AF}_1$, the related admissible set for argument A is $\{A\}$ since no other argument attacks or defends A. Consider $\mathcal{AF}_2$ in Figure 2; this is the same AF as in Figure 1 with an additional argument F. F will be part of admissible extensions since F is admissible with respect to any set. However, since there is no attack relation between F and any other, argument F will not be a member of any related admissible extension except the one for F. Every argument in a related admissible extension is *related to* the topic argument of the extension.

Definition 2.10 (Related to) (Fan & Toni, 2015b, Definition 3.) Given $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$, let $A, B \in \mathcal{A}$. Then A is *related to* B iff:

1. $A = B$; or
2. $(A, B) \in \mathcal{R}$; or
3. $\exists C \in \mathcal{A}$, such that $(A, B) \in \mathcal{R}$ and C is related to B

Informally, all arguments related to an argument can be found by following attack relations backwards in the argument graph.

An *explanation* for argument A is a related admissible extension with A as a topic. There can be many explanations for the same argument. Fan and Toni (2014) define *verbose* and *compact* explanations as different methods for selecting explanations. This formalized the selectiveness property of explanations as described by (Miller, 2019). A set of arguments is *verbose* if it includes as many relevant reasons for explaining an argument as possible. A set of arguments is *compact* if none of the reasons for explaining an argument can be eliminated.

Definition 2.11 (Fan & Toni, 2014, Definition 4.) Given an $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$ and an argument $A \in \mathcal{A}$, let $E_A = \{S \mid S \text{ is a related admissible set with topic } A\}$. Then, for any $S \in E_A$, S is

- A *compact explanation* for A iff S is smallest (w.r.t $\subseteq$) in E_A;
- A *verbose explanation* for A iff S is largest (w.r.t $\subseteq$) in E_A.

Example 2.2 Consider $\mathcal{AF}_3$ in Figure 3. In this AF with topic A, $\{A, D, E, F\}$, $\{A, D, E\}$, $\{A, D, F\}$, $\{A, E, F\}$, and $\{A, E\}$ are related admissible. The set $\{F, G\}$ is admissible but not related admissible because there is no relation between A and G. The verbose explanation for A in this AF is $\{A, D, E, F\}$. There are two compact explanations: $\{A, D, F\}$ and $\{A, E\}$.

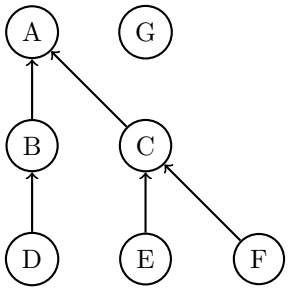


Figure 3: A graphical representation of $\mathcal{AF}_3$ (Example 1 from (Fan & Toni, 2015a)).

In this approach, explanations are fully defined and constructed in abstract argumentation. One danger of directly modelling examples in abstract AFs is that it can lead to ad-hoc modellings instead of utilizing an argument and attack structure. Additionally, it is important to note that abstract accounts of argumentation may make implicit assumptions not shared by many of their instantiations, which could make it challenging to apply and generalize these methods to all instances of abstract argumentation and real-world settings (Caminada & Wu, 2011; Prakken & de Winter, 2018).

The work by Fan and Toni (2014) has been generalized by Borg and Bex (2021a). They define a basic framework for explanations in argumentation that can work for many different types of explanations and with various semantics. This framework works in both abstract and structured argumentation. In the framework, basic explanations are defined in terms of two functions: $\mathbb{D}$, the depth of the explanation and $\mathbb{F}$, the form of the explanation. The form $\mathbb{F}$ determines what part(s) of an argument should be used for explanation. For $\mathbb{F}$, the functions to retrieve components of arguments in section 2.3.2 can be used. The depth $\mathbb{D}$ determines how far away from the topic arguments should be considered for an explanation. This $\mathbb{D}$ can also be used to represent the related semantics from Fan and Toni (2014) using the following definition as proven by Borg and Bex (2021a).

Definition 2.12 Let $\mathcal{AF} = (\mathcal{A}, \mathcal{R})$, $A \in \mathcal{A}$, and $\mathcal{E}$ is an extension of $\mathcal{AF}$ for some semantics.

- $\texttt{DefBy}(A) = \{B \in \mathcal{A} \mid B \text{ defend } A\}$;
- $\texttt{DefBy}(A, \mathcal{E}) = \texttt{DefBy}(A) \cap \mathcal{E}$, the set of arguments in $\mathcal{E}$ that defends A;
- Then $\{\texttt{DefBy}(A, \mathcal{E}) \mid \mathcal{E}$ is an admissible extensions of $\mathcal{AF}$ which contains $A\}$ is the set of all related admissible explanations for A.

3 Experiments in Argumentation

Several methods for generating explanations based on argumentation frameworks were introduced in the previous section. In these methods, arguments are selected based on criteria that supposedly align with human cognitive biases such as causality and selectiveness (Miller, 2019). While this suggests that these explanation methods are cognitively plausible, empirical evaluation is lacking. Several studies have evaluated the correspondence between different aspects of formal argumentation and human cognition and reasoning, such as rebuttal (Definition 2.7), semantics (Definition 2.1), and reinstatement (defending an attacked argument). This section will provide an overview of these studies. These studies can be categorised into *framework-first methods*, which start with an AF and perform experiments based on this, and *dialogue-first methods*, which begin with free-form, semi-structured or structured dialogue and evaluate properties of formal argumentation based on this. An overview of all framework-first methods can be found in Table 2, and an overview of all dialogue-first methods can be found in Table 3.

3.1 Framework-first methods

The first major experiment investigating the cognitive plausibility of formal argumentation was performed by Rahwan et al. (2010). They performed two experiments, one concerning simple and one concerning floating reinstatement. The arguments in the experiments were based on the AFs in a) and b) in Figure 4. In the simple reinstatement case, argument A is defeated by argument B, which is, in turn, defeated by argument C. In this case, C reinstates A. In the floating reinstatement case, both C and D reinstate argument B, but C and D attack each other. In both experiments, participants assessed the conclusion of arguments on a 7-point Likert scale. Participants were shown natural language arguments based on these AFs; no explicit attack relations were shown. In the simple reinstatement experiment, participants were first shown a base argument (A), then a defeater (B) was added, and finally, a third argument (C) was added, which reinstated the initial argument (A). An example of these arguments can be found in Example 3.1. This was repeated for six different sets of contents, leading to 18 different assessments being collected per participant.

Example 3.1 Argument set 1 from experiment 1 (Rahwan et al., 2010).

(A) The battery of Alex's car is not working. Therefore, Alex's car will halt.
(B) The battery of Alex's car has just been changed today. Therefore, the battery of Alex's car is working.
(C) The garage was closed today. Therefore, the battery of Alex's car has not been changed today.

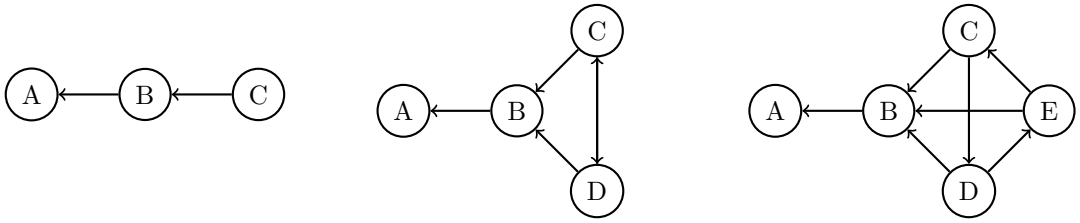
The floating reinstatement experiment used the same experimental set-up, but instead of being presented with one additional argument in the reinstatement case, two arguments (C & D) were presented that attack each other and the second argument. One of the argument sets used can be found in Example 3.2.

Example 3.2 Argument set 1 from experiment 2 (Rahwan et al., 2010).

(A) Cody does not fly. Therefore, Cody is unable to escape by flying.
(B) Cody is a bird. Therefore, Cody flies.
(C) Cody is a rabbit. Therefore, Cody is not a bird.
(D) Cody is a cat. Therefore, Cody is not a bird.

Results were the same in both experiments. Confidence in the argument was the highest in the base case, lower in the defeated case, and back up (but not fully restored) in the reinstated case. The lack of a full recovery of arguments is relevant because formal semantics do not predict this. In the second experiment, argument A was accepted by participants after floating reinstatement; this is evidence of a preference for preferred over grounded semantics. Based on these experiments Rahwan et al. (2010) advise avoiding discussion or explanation that might reveal a defeater to an argument since this reduces confidence. However, if a defeater is revealed, confidence can be partially restored by defending the argument.

Rahwan et al. (2010) speculate that the absence of full reinstatement could be caused by a disruption in the 'suspension of disbelief'. Because participants have heard counterarguments to an argument, even



(a) Simple Reinstatement (b) Floating Reinstatement (c) 3-cycle Reinstatement

Figure 4: The AFs in a) and b) are used by Rahwan et al. (2010) and Bezou Vrakatseli et al. (2021), the AF in c) was added by Cramer and Guillaume (2018a) and used by Guillaume et al. (2022).

though these arguments are refuted, they might be more inclined to think critically about the topic argument. In their 2021 study Bezou Vrakatseli et al. tested different methods of presenting arguments to see if this would change the results, using three experiments. In the first experiment, they replicated Rahwan et al. (2010)'s findings. In the second experiment, all natural language arguments were presented simultaneously, meaning participants saw the entirety of Example 3.1 at once. The third experiment presented the general theory from which specific arguments were drawn before the arguments were shown. A general version of the arguments in Example 3.1 can be found in Example 3.3.

Example 3.3 A generalized version of Example 3.1.

- A car will halt if its battery is not working.
- A car's battery is working if it has been changed the same day.
- When the garage is closed, a car's battery cannot be changed.

In all three experiments, 130 participants rated eight sets of three simple reinstatement arguments. They rated the acceptability of the topic argument on a 7-point Likert scale. Results from the second experiment were the same as in the first experiment, and the results found by Rahwan et al. (2010). In the third experiment, the base confidence was lower, and the confidence in the reinstatement condition was higher than in the other conditions. Bezou Vrakatseli et al. speculate that the base confidence is lower because participants have seen attackers that have not been ruled out. Confidence after reinstatement is higher because all presented attackers have been defeated. Neither experiment managed to find full reinstatement. Thus, these experiments support Rahwan et al.'s findings that full reinstatement does not align with how humans evaluate arguments.

In these studies by Rahwan et al. (2010) and Bezou Vrakatseli et al. (2021), abstract argumentation frameworks were instantiated using natural language. In these AFs, conflicts between arguments are presented in a non-symmetric directed manner. Cramer and Guillaume (2018a) investigated if the way people interpret the conflict between arguments corresponds to this. They performed two experiments. For the first experiment, they recruited 'naive' adults with no or limited knowledge of argumentation. For the second experiment, they recruited experts in formal argumentation. Both studies used AFs with three different structures: simple reinstatement, floating reinstatement, and 3-cycle reinstatement. These can be found in Figure 4. These first two AFs correspond to the ones used by Rahwan et al. and Bezou Vrakatseli et al.. These AFs were filled with natural language arguments from various contexts to create 40 argument sets. The 'naive' participants were asked to judge the acceptability of pairs of arguments. Experts were shown the full argument set and asked to indicate all attack relations. Cramer and Guillaume (2018a) found similar results in both studies and concluded that conflicting arguments can be created that people interpret as unidirectional attacks. They also found that undercutting the trustworthiness of a source was the most convincing unidirectional attack.

These AFs validated by Cramer and Guillaume (2018a) have been used by Guillaume et al. (2022) to test the cognitive plausibility of various semantics. In the AFs for floating and 3-cycle reinstatement, semantics disagree on the acceptability of arguments. In the first experiment, to verify that attack relations were seen in accordance with the intended AFs, participants were, in groups, tasked with drawing all attack relations between arguments in a set. In the second experiment, participants evaluated

the acceptability of these arguments. Participants first wrote down an initial answer and then discussed their answers in a group to improve logical reasoning, after which participants individually wrote down their final answer. For each of the three AFs, accuracies increased after group deliberation. Participants drew the correct attack relations for simple reinstatement. For floating reinstatement, participants mostly agreed with the AF. However, there was no consensus on the bilateral attack between C and D. In 3-cycle reinstatement, participants mostly agreed with the AF, except there was no significant agreement with the attacks from E to C and from D to B. For floating reinstatement, participants' judgements aligned with preferred and CF2 semantics but not grounded semantics since argument A was accepted. For 3-cycle reinstatement, CF2 was closest aligned to the participants' judgements since participants did not defend all arguments in the accepted set. Agreement differed slightly between the contexts used for the arguments, possibly because world knowledge influenced judgements.

Similar results were found by Cerutti, Tintarev, and Oren (2014). In their experiment, participants were presented with natural language arguments from four different contexts and were tasked with indicating their agreement with the conclusions of the arguments. Participants also indicated why they chose this option and how relevant other arguments were. They found that preference between arguments is domain-dependent, and participants justify their choice with domain-specific reasons, such as *"All weather forecasts are notoriously inaccurate"*. Hadoux & Hunter, 2019 also tested participants' preferences in persuasive dialogues. In their experiment, participants were first explicitly asked about their preferences and possible reasons for choosing one argument over another. After this, they were presented with sets of arguments and asked to indicate their preference. Participants expressed preferences lined up with their behaviour, and there was moderate agreement on preferences between the participants. Knowledge of these preferences can be used to increase the persuasiveness of dialogues, especially if domain-specific preferences are taken into account.

In a follow-up study, Cramer and Guillaume (2019) attempted to eliminate this effect of world knowledge by instantiating their AFs with arguments relating to a hidden treasure on a remote island. They used 12 sets of natural language arguments corresponding to 12 argumentation frameworks chosen to highlight differences between semantics. All attacks were based on the information that a source is not trustworthy since this was found to be the most convincing unidirectional attack by Cramer and Guillaume. Participants were presented with the natural language arguments as well as a graphical visualization of the argumentation framework and asked to indicate the status of these arguments on a three-point scale (accepted, rejected, undecided). The experiment also included a group deliberation phase to improve logical reasoning. The acceptance status of arguments in grounded, preferred, semi-stable, CF2, stage, and stage2 semantics were compared to human judgements. CF2 and grounded semantics were the best predictors of human argument evaluation. This is noteworthy because Guillaume et al. (2022) found preferred semantics to be a better predictor than grounded semantics. The authors speculate that participants gravitate towards grounded semantics in more complex scenarios. Thus, in this experiment, grounded semantics predicted the cognitively simpler strategy of choosing 'undecided' when there is doubt, and CF2 semantics predicted more complex strategies for determining acceptability.

Yu et al. (2018) tested whether participants in their experiment preferred a restricted or unrestricted rebut. The latter allows rebuttal on all arguments that contain at least one defeasible rule, whereas the former only allows rebuttal if the most recent rule is defeasible. In their experiment, participants were asked in a survey whether they felt presented counterarguments were legitimate responses to presented arguments and if the counterargument actually attacked the argument. They found participants' responses are more in line with unrestricted than restricted rebuttals.

Polberg and Hunter (2018) used probabilistic argumentation to represent the extent to which an argument is believed or disbelieved. In their experiment, participants were presented with arguments from a dialogue. After every argument, they indicated their agreement with the argument and explained their agreement. Participants were also asked how the presented argument related to other arguments. They found that different participants interpret statements and relations between them differently and that their knowledge can affect this. The authors also found that even when participants are sure two arguments are connected, they might not be sure of the relationship between these arguments. Polberg and Hunter also question whether three values (accepted, rejected, undecided) are enough to capture

the participant's judgement of arguments. They also found that the notion of defence does not account for all positive relations between arguments since some relations are explicitly described as supporting. Based on these findings, the authors conclude that the most common approaches in argumentation might be too simplistic to grasp human reasoning, and probabilistic and bipolar frameworks might be more cognitively plausible.

Study	Topic	FA	Methods	Outcome
Rahwan et al., 2010	Reinstatement	A	Participants rate confidence after introduction of arguments	Confidence drops after defeat and is partially restored after reinstatement
Bezou Vrakatseli et al., 2021	Reinstatement	A	Replication of Rahwan et al., present arguments all at once, first present all general scenarios.	Findings of Rahwan et al. were replicated in all three settings
Cramer & Guillaume, 2018a	Directionality of attacks	A/S	Naive participants and argumentation experts indicate acceptability status of arguments.	Conflicting arguments that people interpret as unidirectional attacks can be created.
Guillaume et al., 2022	Semantics	A/S	Participants draw all attack relations and judge acceptability in an AF.	Preferred and CF2 semantics are better predictors than grounded and complete semantics.
Cramer & Guillaume, 2019	Semantics	A	Participants evaluate the acceptability status of natural language arguments.	Grounded and CF2 performed equally well. Grounded as simple, CF2 as a complex strategy.
Cerutti et al., 2014	Argument preference	S	Assess preference between two conflicting Natural language arguments	Preference relations are domain-dependent
Hadoux & Hunter, 2019	Argument preferences	A	Participants' stated preferences were compared to actual argument preferences	agreement on preferences between participants and within participants' preferences.
Yu et al., 2018	Rebuttal	S	Assess the strength of counterarguments and if the counterargument actually attacked arguments.	Responses are more in line with unrestricted than restricted rebuttal.
Polberg & Hunter, 2018	Uncertainty of relations	A	Participants rate and describe relations of arguments in dialogue	Common approaches in argumentation might be too simplistic to grasp human reasoning

Table 2: An overview of framework-first experiments, FA stands for Formal Argumentation and indicates whether experiments were performed using (A)bstract or (S)tructured argumentation frameworks.

3.2 Dialogue-first methods

The previously discussed studies start with an argumentation framework and experimentally test aspects of these frameworks. Several studies have been conducted using a more inductive approach to determining AFs. One such study by Rosenfeld and Kraus (2014) used argumentative conversations from the Penn Treebank Corpus (1995). The authors manually constructed bipolar argumentation frameworks (BAF) based on these dialogues. Dialogues not used to construct the BAFs were then annotated using the BAFs. They then tested several concepts in formal argumentation using these frameworks to see if people followed them. Conflict-freeness, expected in all argument sets, was only present in 78 per cent of arguments. Scores were even lower for acceptability, admissibility, and arguments being part of extensions. Rosenfeld and Kraus conclude that formal argumentation is not predictive or descriptive of free-form human argumentation.

In a 2016 study, Rosenfeld and Kraus further expand on this with two experiments using semi-structured conversation. In the first experiment, participants are presented with an existing conversation and tasked to choose the next response out of four options. In the second experiment, participants took part in semi-structured online chats in which participants could choose from a predefined list of arguments along with some extra conversational options such as 'however', 'but', and 'additionally'. Grounded, preferred and stable extensions were calculated for all AFs in these experiments. People frequently chose non-justified arguments, and the acceptance status of arguments based on these extensions was a poor predictor of participants' responses. Rosenfeld and Kraus introduce a *relevance* heuristic as a simple prediction method. Where relevance was defined as both the path length from the currently presented argument to the latest argument and from the current argument to the topic argument. This heuristic can be used to select relevant arguments just like the notion of $\mathbb{D}$ in (Borg & Bex, 2021a). This heuristic provided a better prediction of human behaviour than any of the argumentation semantics in the study.

Villata et al. (2017) investigated the connection between argumentation, emotion, and personality in online debate interactions. They performed an experiment in which participants debated topics while their emotional state was measured using webcams and an EEG headset. Their personalities were also assessed. Villata et al. took into account whether arguments brought forward by participants were supporting or attacking the topic statement. The researchers manually annotated this. Out of 6 basic emotions, anger and disgust were most frequently measured. Increased anger was correlated with increased engagement. This might mean that when participants' arguments are refuted, they spend more attention coming up with counterarguments. Participants also provide more arguments as their disgust increases and fewer arguments when sadness is detected. Participants who tend to attack others appear less angry than those who attack fewer arguments. Personality also affects the expression and debating behaviour: extroverts showed more facial expression, non-conscientious people had a lower cognitive workload, and anxious participants were less engaged. A contradiction on an in-depth conviction provokes strong emotions, which affects the number of arguments participants produce.

Study	Topic	FA	Methods	Outcome
Rosenfeld & Kraus, 2014	Argumentative behavior	A	Annotated free-form conversations compared to argumentation requirements	Conflict-freeness only present in 78 percent. Even lower for acceptability, admissibility, and arguments being part of extensions.
Rosenfeld & Kraus, 2016	Argumentative behaviour	A	Participants chose the next responses to an existing conversation and did semi-structured online chats.	People frequently choose non-justified arguments, and semantics were poor predictors of responses
Villata et al., 2017	Emotions and personality traits	A	Participants debate about topics with either support or attack arguments	Anger and disgust were most frequently measured

Table 3: An overview of dialogue-first experiments, FA stands for Formal Argumentation and indicates whether experiments were performed using (A)bstract or (S)tructured argumentation frameworks.

4 Empirical Study

4.1 Research question

Several methods for selecting explanations for a topic argument based on an argumentation framework have been discussed in Section 2.3.3. One recurring aspect present in several theories of explanation is *relevance*. Relevance is important to the definition by Verma et al. (2020) of explanation in Section 2.2, where relevance is used in *the transfer of relevant knowledge* between the explainer and the recipient of an explanation. The relevance of elements in an explanation is also vital for some of the desirable features of explanations which were discussed in Section 2.2.2. When providing an explanation, people only give a small subset of possible causes. This subset of causes should be selected to be relevant to the recipient (Miller, 2019). Relevance was also discussed in relation to explanations based on formal argumentation. *Related admissible* extensions (Definition 2.9) only include arguments relevant to the topic argument by only including arguments that defend it. A related admissible extension is an explanation for the topic argument of that extension. Individual arguments are related to themselves and to every argument they directly or indirectly attack (Definition 2.10). In Section 2.3.3, the *relevance heuristic* by Rosenfeld and Kraus (2016) was discussed. In this heuristic, an argument's relevance is defined by the path length from the currently presented argument to the latest argument and from the current argument to the topic argument. Each of these studies defines a notion of relevance in formal argumentation to provide better explanations than can be done with standard argumentation semantics by formalizing human preferences and biases. However, no existing studies have investigated whether these proposed explanation methods align with human explanation behaviour. Therefore, the research question for this study is:

Research question: Are explanations based on relevance in formal argumentation cognitively plausible?

To be cognitively plausible, a system must perform (roughly) the same as humans do on a cognitive task or be plausibly built on components that have met this test (Kennedy, 2009). To answer the research question, explanations provided by people need to be compared to explanations based on relevance in formal argumentation. This research question will be answered using two sub-questions, each testing different notions of relevance.

Sub-question 1: Do people include unrelated arguments in explanations?

This question aims to test the notions of relatedness (Fan & Toni, 2015a) and related admissible semantics (Fan & Toni, 2014) by investigating whether people include arguments in an explanation that are not connected to the topic argument. Based on the desirable features of explanations (Section 2.2.2), people should prioritize relevant information and thus not include unrelated arguments in explanations. This would provide empirical evidence for related admissible semantics over other formal argumentation semantics and for explanations based on relevance in a broader sense.

Sub-question 2: Do people prefer direct over indirect defenders?

This sub-question focuses on the relevance heuristic by Rosenfeld and Kraus (2016). According to this heuristic, people prefer arguments with a shorter distance to a topic argument over arguments that are further away. Direct defenders are closer to the topic argument than indirect defenders and are thus expected to be preferred in explanations over indirect defenders. In the definitions of relatedness and relevance by notions of relatedness and relevance by Fan and Toni (2014) and Borg and Bex (2021b), no distinction is made between direct and indirect defenders. A positive answer to this sub-question could indicate that these definitions are not cognitively plausible because they are not selective enough in their explanation methods.

4.2 Methods

In the following section, the methodology used to investigate the research question will be discussed. First, the argumentation frameworks used to study each sub-question will be introduced. Next, these AFs were instantiated using natural language arguments. The instantiating of the abstract AFs with natural language arguments was validated using a pilot study. In this pilot study, it was tested whether the AFs were interpreted by participants as intended. Based on this pilot study, the most appropriate argument sets for the main study were selected. Finally, this section will outline the procedure and the analysis of the results of the empirical study.

4.2.1 Argumentation frameworks

Previous research into formal argumentation using human participants has focused on asking about the acceptability of individual topic arguments (Cramer & Guillaume, 2018a, 2019; Guillaume et al., 2022) or on having participants rate topic arguments on a 7-point scale from *certainly false* to *certainly true* (Bezou Vrakatseli et al., 2021; Rahwan et al., 2010). It is not possible to take a similar approach to evaluate explanations for two reasons. Firstly, an explanation always consists of a set of one or multiple arguments in relation to a topic. Thus, no one argument can be individually evaluated. The topic argument can be evaluated based on a shown explanation. However, such a setup would not measure how people explain but how acceptance of a topic changes based on an explanation. This is unsuitable for the current research question because people's behaviour needs to be observed to investigate cognitive plausibility. The second difficulty of evaluating explanations is that a defender only makes sense as an explanation for a topic argument in relation to the attacker that is being defended from. In Example 3.1, argument C: *" The car had just undergone maintenance service. Therefore, the brake fluid was not empty."* only makes sense as a defender for argument A: *"Louis applied the brake and the brake was not faulty. Therefore, the car slowed down."* knowing that A is attacked by B: *"The brake fluid was empty. Therefore, the brake was faulty."* Explanations cannot be presented without the required context. Therefore, this study used a different methodology. Participants were presented with a topic argument at the top of the page and all the natural language arguments in an AF in random order below the topic. Arguments were shown in random order to make sure participants had to read each argument set carefully and could not select the same explanation every time. After this, they were tasked to choose those arguments that they believed to be an appropriate explanation for the conclusion of the topic argument.

Each sub-question has been investigated using a different AF. The first sub-question of this study is, *"Do people include unrelated arguments in explanations?"*. The simplest AF that can be used to investigate this sub-question can be seen in Figure 5. This AF is suitable for this question because it includes an unrelated argument, E, and an argument that can explain the topic, C. This AF has both a related admissible extension and an unrelated admissible extension. The admissible but unrelated explanation for A is $\{C, E\}$. The admissible related explanation for A is $\{C\}$.

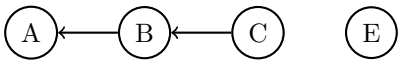


Figure 5: $\mathcal{AF}_4$ for hypothesis 1.

For the second sub-question, *"Do people prefer direct over indirect defenders?"*, direct and indirect defenders were compared. For this $\mathcal{AF}_5$ in Figure 6 was used. The direct defender of A is C, and the indirect defender is E.

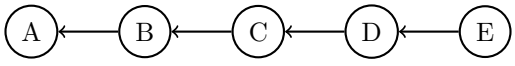


Figure 6: $\mathcal{AF}_5$ for hypothesis 2.

4.2.2 Arguments

In the study, participants were not shown the abstract representation of the AFs. Instead, they were shown natural language instantiations of the AFs. Instantiating AFs using natural language arguments

is not a trivial task. First, context can influence the interpretation of an AF. What people already know about a topic changes how they see arguments (Cerutti et al., 2014; Cramer & Guillaume, 2018a). And second, people might not interpret the AF as the intended structure. This can be an issue when constructing natural language-based AFs from scratch without an underlying system that arguments are drawn from (Caminada & Wu, 2011; Prakken & de Winter, 2018). Therefore, in this study, arguments validated in previous research have been used alongside the arguments created for this study. All argument sets have been validated using a pilot study. In these argument sets, all attack relations are undermining attacks, where the conclusion of an argument negates the premise of another argument. Since the structure of the AFs in previous studies differs from those in this study, the argument sets needed some modifications before they could be used. Unrelated arguments had to be developed for this study because no previous study has used AFs with them. Unrelated arguments were created not to attack and not to be attacked by any existing argument. As an additional requirement, unrelated arguments were set in the same context as the other arguments in the set, with the aim of not making the 'unrelatedness' trivial. This was deemed more realistic since, in an XAI application, all arguments would come from the same system and thus be in the same context. The complete set of natural language arguments for $\mathcal{AF}_4$ can be found in Appendix B, and one example can be found below. To instantiate $\mathcal{AF}_5$, arguments D and E needed to be added to some existing structures. From other structures, arguments were removed to create $\mathcal{AF}_5$. The complete set of natural language arguments for this AF can be found in Appendix D, and one example can be found in Example 4.2.

Example 4.1 (Arguments for $\mathcal{AF}_4$) *A would be the topic argument for explanations.*

(A) Stephen is not guilty. Therefore, Stephen is to be free from conviction.
(B) Stephen was seen at the crime scene at the time of the crime. Therefore, Stephen is guilty.
(C) Stephen was having dinner with his family at the time of the crime. Therefore, Stephen was not seen at the crime scene at the time of the crime.
(E) Stephen is very tall. Therefore, Stephen likely doesn't leave small footprints.

Argument E is unrelated to the other three arguments since there is no formal attack to or from E. However, since argument E mentions Stephen and footprints, it is set within the same context of a criminal investigation relating to Stephen.

Example 4.2 (Arguments for $\mathcal{AF}_5$) *A would be the topic argument for explanations.*

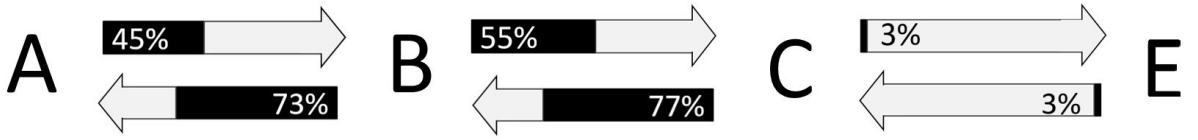
(A) The battery of Alex's car is not working. Therefore, Alex's car will halt.
(B) The battery of Alex's car has just been changed today. Therefore, the battery of Alex's car is working.
(C) The garage was closed today. Therefore, the battery of Alex's car has not been changed today.
(D) Alex works at the garage. Therefore, Alex can change the battery of their car when the garage is closed
(E) Alex just got a new job. Therefore, Alex does not work at the garage anymore.

4.2.3 Pilot study

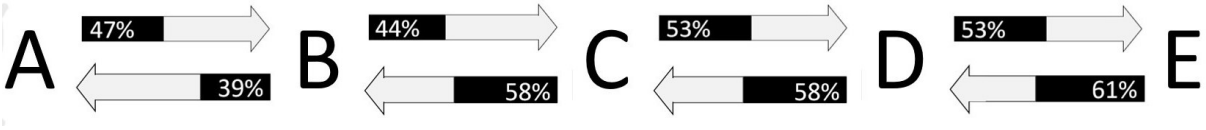
A pilot study was conducted to validate the instantiation of the AFs using natural language arguments as described above. The pilot also aimed to test whether the sentences were easy to read even for non-native speakers and the comprehension of arguments when presented in random order. The latter was relevant since previous studies from which arguments have been used presented arguments in a set order. In this pilot, five participants were presented with sets of arguments and asked to indicate the attacks between arguments. Then, these indicated attacks were compared to the intended abstract structure. Argument sets for which participants agreed with the intended AFs were kept for the study. Two different instruction methods were tested; the attacks indicated by participants were more in line with the intended AFs when they received a more detailed explanation of the definitions of 'argument' and 'attack'. The summary of all attacks drawn by participants in the pilot can be found in Figure 7.

For $\mathcal{AF}_4$, the most frequently indicated attacks were the intended attacks, and barely any attacks were indicated towards or from the 'unrelated' argument E. When given more detailed instructions, $\mathcal{AF}_5$ was interpreted as intended in most cases. Based on the pilot results, six argument sets for $\mathcal{AF}_4$ and

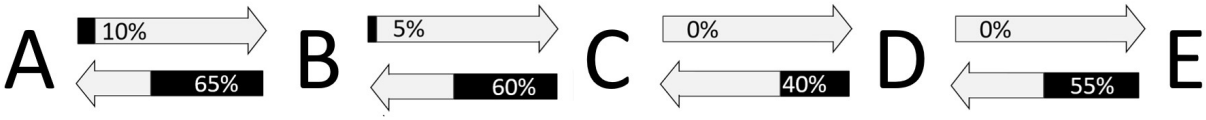
six sets for $\mathcal{AF}_5$ were selected for the study. The selected argument sets can be found in Appendix C. Wording changes were made based on input from pilot participants to make arguments easier to read and to reduce ambiguity. This pilot study's entire methods and results can be found in Appendix A.



(a) $\mathcal{AF}_4$ with standard instructions



(b) $\mathcal{AF}_5$ with standard instructions



(c) $\mathcal{AF}_5$ with detailed instructions

Figure 7: Percentage that each attack between adjacent arguments was indicated in the pilot for each of the two AFs. Some attacks between non-adjacent arguments were also drawn but are not shown in this figure.

4.2.4 Participants

For this study, 127 participants were recruited through convenience sampling by asking friends and family and using social media. Anyone above 18 could participate in the study. Most participants (80) were between 18 and 35 years old. The other 47 participants were older than 35. Since this study was conducted in the Netherlands, most participants were not native English speakers. Therefore, participants were asked about their English reading proficiency at the end of the survey. Participants rated their English reading proficiency as 4.2 out of 5 ($SD = 0.75$). Most participants (40) obtained a bachelor's degree, but the sample included almost an equal amount of participants with a master's degree (39) or PhD (35). For eight, high school was their highest completed level of education, and three completed vocational education. Two participants chose not to answer this question. Participants' prior knowledge of formal argumentation might have influenced their behaviour in the study. Therefore, participants were asked about their familiarity with formal argumentation at the end of the study. About half of the participants (63) reported being unfamiliar with or never having heard of formal argumentation. Of the other half of the participants, 31 were somewhat familiar, and 33 reported being very familiar with or experts in the field. All participants gave informed consent, and the Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences was conducted (Appendix G). It classified this research as low-risk, with no additional ethics review or privacy assessment required.

4.2.5 Procedure

This study used an observational non-experimental design. Participants' choices of explanations were collected. There was no manipulation of an independent variable, and all participants performed the task under the same conditions. The study was conducted using the online survey tool Qualtrics. Participants received a digital invitation with a link to participate.

At the beginning of the study, participants were asked for informed consent (Appendix E), and the procedure of the study was explained (Appendix F). Participants were presented with eight sets

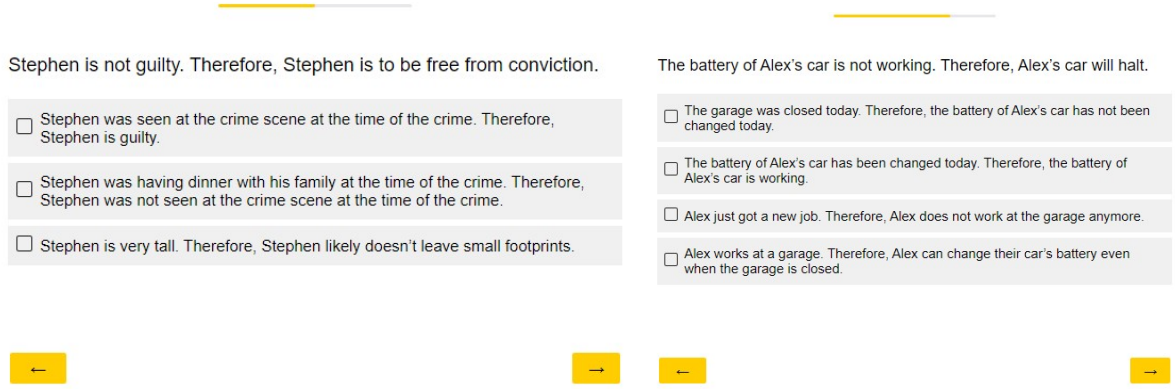


Figure 8: Examples of arguments for the two different AFs as presented in the online survey.

of arguments in random order. These eight sets were randomly selected from the complete collection of 12 argument sets (Appendix C). Four of these eight sets corresponded to $\mathcal{AF}_4$, and the other four corresponded to $\mathcal{AF}_5$. Participants were not informed of the underlying AFs and were only presented with the natural language arguments that instantiated the AF. In a single trial, participants were shown a topic argument and three or four other arguments (depending on the AF). These arguments were shown in random order, except for the topic argument, which was always first. Then participants were asked to select the argument(s) that, according to them, explain(s) why the conclusion of the topic argument is the case. Participants could choose as many arguments as they wanted and had to select at least one argument. Figure 8 shows an example of the layout.

After providing an explanation for eight arguments, participants were asked four demographic questions and had the option to leave a comment on the survey. First, they were asked to indicate their age using five bins (18-25, 26-35, 35-45, 45-65, 65+). Second, they were asked to rate their English reading ability on a 5-point scale from 'extremely bad' to 'extremely good'. Third, participants were asked about their highest completed level of education. Finally, they were asked to indicate their familiarity with formal argumentation based on five options ('I have not heard of it', 'I'm not familiar with it', 'I've some familiarity (ex. had one course on it)', 'I'm quite familiar with it', and 'I'm an expert'). The study's layout, including these four questions, can be found in Appendix G.

4.2.6 Analysis

After data collection, all data was anonymised for analysis by removing response IDs from the data set. Incomplete responses were eliminated, and responses were checked for anomalies, such as giving the same answer for each question. No such anomalies were found. Then, responses were aggregated over all participants for each argument set and analysed in four steps. These steps are briefly outlined here and illustrated in Figure 9. First, tests were conducted to confirm whether argument sets per AF were answered consistently. Second, the distribution of explanations per AF was compared to chance to determine if there was a pattern in answers that could be explored. Third, the evidence for the hypotheses was weighed against the evidence against them. Finally, individual answer proportions were compared to determine which explanations provided the most support for and against the hypotheses. Further details will be provided below.

Consistency of answers for the two different AFs was tested in the first analysis step. In this study, six natural language argument sets were used for each AF. All argument sets that instantiate the same AF was intended to be interpreted the same by participants. Differences in answer behaviour between argument sets would indicate that participants did not interpret all argument sets for one AF the same. If this were to be the case, then explanations collected for each argument sets should not be merged for further analysis but instead analysed separately. Two chi-square tests for independence were used to assess the consistency of the different natural language argument sets. This test is significant if a difference exists between the answers selected for the natural language argument sets. One test was

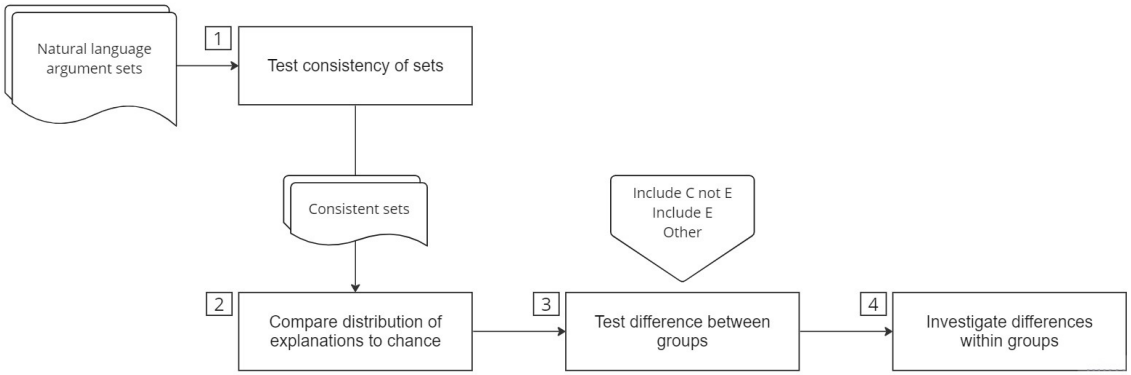


Figure 9: A graphical overview of the steps taken in the analysis.

conducted per AF. All argument sets for $\mathcal{AF}_4$ were compared to each other, and all sets for $\mathcal{AF}_5$ were compared to each other. If the test was significant, there is a relationship between the given explanation and the sets of arguments; given explanations are different for different argument sets. If this were to be the case, there are some sets for which the given explanations differed from others. To identify these sets, individual chi-square tests were performed, with which each set was compared to the average distribution over all six sets. The sets that deviated were investigated to determine whether there was a logical or theoretical reason for the deviation. Based on this investigation, the sets were included in the further analysis, separately analysed, or excluded from further analysis.

Selected explanations were compared to chance in the second analysis step. The goal of this step was to determine if there was a pattern in answers that could be further explored. If the explanations given by participants were totally random, it would be not very sensible to take any further analysis steps. For this purpose, a chi-square goodness-of-fit test was conducted for each AF, which compared the frequencies with which all explanations were selected to the expected frequencies if all answer options were selected equally. If a significant result were to be found from this test, there is a pattern in answers that can be further explored, which will be done using the following analysis steps.

The difference between three groups of answer options was tested to answer each sub-question. The first sub-question of this study is, *"Do people include unrelated arguments in explanations?"*. If people are indeed selective in choosing explanations as discussed in Section 2.2.2, then it would follow that participants' responses only include arguments related to the topic argument. For the argument sets for $\mathcal{AF}_4$, participants were expected to never select argument E as (part of) an explanation and always select argument C. This was expected because E is unrelated to the topic, and C is the direct defender and related admissible explanation of the topic. For the first sub-question, participants selecting explanations $\{C\}$, $\{B, C\}$ would support and all explanations, including E ($\{E\}$, $\{B, E\}$, $\{C, E\}$, $\{B, C, E\}$) would undermine the hypothesis that people do not include unrelated arguments in the hypothesis. Participants may also select neither C nor E, which would be explanation $\{B\}$.

For the second sub-question, *"Do people prefer direct over indirect defenders?"*, it was hypothesised that participants would prefer direct defenders over indirect defenders. In the AF for this sub-question argument, C is the direct defender, and E is the indirect defender. Thus, explanations containing C but not E were expected to be selected more than explanations containing E. This hypothesis would be supported if the proportion $\{C\}$ and $\{B, C\}$ explanations given was larger than the proportion of $\{E\}$, $\{B, E\}$, $\{C, E\}$ and $\{B, C, E\}$ explanations. All other possible explanations ($\{B\}$, $\{D\}$, $\{B, D\}$, $\{D, E\}$, $\{C, D\}$, $\{B, D, E\}$, $\{B, C, D\}$, $\{C, D, E\}$ $\{B, C, D, E\}$) were considered to neither support nor undermine the hypothesis since they either included neither C nor E because they were nonsensical explanations indicating that participants possible did not understand the task. Explanations were deemed nonsensical if they included argument D since this argument is not part of any acceptable extension and does not provide context to connect an argument to the topic, such as argument B does for C.

For each AF, the proportions of answers that fell into each of the three groups were calculated and compared using two-sample tests for equality of proportions. If a significant result was found, the differences in proportions within groups were investigated to determine which explanations were the major contributors to this significant result.

Within group differences were investigated to identify which individual answers contributed most to the difference found in the previous step. For this purpose, the proportions with which each explanation was chosen were calculated and compared using two-sample tests for equality of proportions. Significant differences between proportions within groups of explanations will show which individual explanations were the drivers of differences between groups.

The primary research question of this study is *"Are explanations based on relevance in formal argumentation cognitively plausible?"*. A significant preference for explanations containing C over all other arguments in both $\mathcal{AF}_4$ and $\mathcal{AF}_5$ would be evidence for the hypothesis that explanations based on relevance in formal argumentation are cognitively plausible. If, in a large proportion of cases, argument C was not selected for $\mathcal{AF}_4$ and neither C nor E was selected for $\mathcal{AF}_5$, responses would not align with any formal semantics, thus it would be impossible to decide between standard and related semantics based on the found results. Such data could be found if participants generate explanations that do not align with formal argumentation or disagree with the intended AF structures. The latter, however, should be ruled out because the natural language arguments for the AFs were validated in the pilot study.

Participants could select arguments B in $\mathcal{AF}_4$ and B and D in $\mathcal{AF}_5$. These arguments are not part of extensions in any semantics and thus were not expected to be part of any explanations. However, participants might have selected these arguments together with other arguments to provide context to the explanation; argument C only defends A because argument B exists as an attacker of A. If participants chose B and D without choosing C or E, this might indicate that participants did not understand the task or did not interpret the attack relations as intended. The frequency with which B and D were selected as standalone answers and as part of larger explanations was investigated.

Further analysis was done by investigating if answers differed based on English proficiency and familiarity with formal argumentation. The purpose of investigating differences based on this demographic information was to determine if some participants might not have fully understood the task or the argument sets.

5 Results

In this section, the study's main results are presented in relation to the primary research question and two sub-questions. The primary goal of this study is to investigate whether explanations based on relevance in formal argumentation are cognitively plausible. Each of the two sub-questions focused on a different aspect of relevance. The hypothesis for the first sub-question is that participants will not include unrelated arguments in explanations. The hypothesis for the second sub-question that that participants chose direct over indirect defenders in explanations. Explanations for eight out of twelve different scenarios by 127 participants were collected. Data was only collected from participants who finished the entire online survey. There were no abnormalities in the data, such as participants taking less than five minutes or choosing the same option for every argument set. Therefore no data was removed.

5.1 Consistency

The first step in the analysis was the asses the consistency of the different natural language argument sets used to instantiate each AF. The consistency of explanations between argument sets indicates whether the argument sets can reliably be combined. In theory, the answers for all natural language argument sets for one AF should be able to be combined since participants should give the same answer for each set. However, if some argument sets are explained differently, participants might not have interpreted all argument sets the same. This should be investigated before aggregation answers over all argument sets. Individual argument sets were grouped by AF and inspected to identify differences between explanation behaviour for the different sets of arguments using a chi-square test for independence.

5.1.1 Argument sets for $\mathcal{AF}_4$

As can be seen in Figure 10, there appears to be a relationship between the argument sets and chosen explanations for $\mathcal{AF}_4$. This is especially noticeable in the number of times $\{C\}$ was selected for each argument set. As seen in Table 4, $\{C\}$ was selected twice as often in set 5 compared to set 4. A chi-square test of independence was performed to assess the relationship between the argument set and chosen explanation. There was a significant relationship between the two variables ($\chi^2(30) = 181.85$, $p < .001$). Therefore, the explanations given for each of the six sets of arguments for $\mathcal{AF}_4$ were individually compared to the total distribution of given explanations using chi-square tests of independence (Table 4). Significant results were found for set 2 ($\chi^2(6) = 30.34$, $p < .001$), set 4 ($\chi^2(6) = 45.82$, $p < .001$), and set 5 ($\chi^2(6) = 18.36$, $p = .005$). The Pearson residuals of the chi-square test and Figure 10 were used to investigate what explanation options contributed most to the difference between these sets and the average distribution. For set 2, $\{C, E\}$ was selected more and $\{C\}$ was selected less than in most other sets. Since $\{C, E\}$ is an admissible explanation for the topic argument, this explanation was not considered non-sensical. Thus, set 2 was kept for further analysis. For set 4, all options but $\{C\}$ were selected more than in other sets and option $\{C\}$ was selected less; the explanations given for this argument set were closer to a uniform distribution than for any other set. This could mean that participants randomly selected explanations for this set. It is possible that participants did not interpret the attacks in this argument set the same as in other argument sets and thus showed different explanation behaviour. The subject of this argument set was the 'trustworthiness' of the different characters in the set. Likely, undermining a character's trustworthiness in an argument was not seen as an attack on that argument. Many of the options selected by participants for this set do not make theoretical sense based on the underlying AF. Therefore, this argument set was not used in further analysis. For set 5, $\{C\}$ was selected more than expected and $\{E\}$ and $\{C, E\}$ were selected less. This set had a lower χ^2 value, so it deviated less from the average distribution than the other two discussed sets. The observed deviation does not indicate that participants did not understand the task; therefore, this set was included in further analysis.

5.1.2 Argument sets for $\mathcal{AF}_5$

For $\mathcal{AF}_5$, there also appear to be differences in the explanations chosen between argument sets, as is evident in Figure 11. There are more than twice the possible explanation options for $\mathcal{AF}_5$ than for $\mathcal{AF}_4$. Therefore, there were many sparse columns in the contingency table (Table 5), which meant that the assumptions of the chi-square test were violated; thus, this test could not be performed. Instead, Fisher's exact test was used, which is more suitable for sparse data. A statistically significant association

existed between the argument set and chosen explanation (p <.001). Therefore, the explanations given for each of the six individual argument sets for $\mathcal{AF}_5$ were individually compared to the total distribution of given explanations using chi-square tests of independence. All argument sets except set 5 significantly deviated from the total distribution (right column Table 5). The highest χ^2 test scores were found for set 4 ($\chi^2(14) = 61.71$, p <.001) and set 6 ($\chi^2(14) = 65.11$, p <.001). In these argument sets $\{C\}$ was selected less than in other sets, and $\{B\}$, $\{D\}$, $\{E\}$, and $\{B, C\}$ were selected more. Overall for these two sets, explanations are equally divided over single-option explanations indicating that participants might have selected options randomly. This was confirmed based on comments left by participants at the end of the survey in which many participants mentioned finding argument set 4 especially difficult and randomly having chosen explanations for this set. Therefore, these two sets were not included in further analysis. With these two sets removed from the total distribution, the other four sets do not significantly differ from this distribution (see Table 6). And therefore, these four sets were kept for further analysis. Noteworthy is the high preference for $\{C, E\}$ in set 2.

	Selected explanation							Chi-square test	
Set	B	C	E	BC	BE	CE	BCE	χ^2 (df=6)	p-value
1	1	60	1	1	0	13	0	16.81	.010
2	9	39	14	0	0	23	1	30.34	<.001**
3	2	71	4	2	0	3	0	13.95	.030
4	18	35	10	15	1	6	4	45.82	<.001**
5	3	77	2	8	0	2	0	18.36	.005*
6	3	54	19	2	0	4	0	16.25	.012
Total	36	336	50	28	1	51	5		

Table 4: All explanations provided by participants per argument set and total for $\mathcal{AF}_4$. The chi-square test column shows the results of the chi-square tests for independence, comparing sets 1 through 6 to the total distribution. * indicates p <.01, ** indicates p <.001.

	Selected explanation															Chi-Square Test	
Set	B	C	D	E	BC	BD	BE	CD	CE	DE	BCD	BCE	BDE	CDE	BCDE	χ^2 (df=14)	p-value
1	3	77	3	1	0	0	0	3	4	1	1	0	0	2	1	30.74	.004*
2	2	41	3	6	1	0	0	0	22	0	1	0	0	0	1	33.71	.001*
3	5	68	1	0	0	0	0	0	3	0	1	0	0	0	1	30.95	.003*
4	9	16	24	16	1	0	2	1	7	1	1	0	0	0	3	61.71	<.001*
5	15	52	4	2	3	0	0	0	9	0	1	0	0	1	2	11.35	.582
6	16	19	22	2	8	2	0	3	0	0	1	0	1	2	0	65.11	<.001*
Total	50	273	57	27	13	2	2	7	45	2	6	0	1	5	5		

Table 5: All explanations provided by participants per argument set and total for $\mathcal{AF}_5$. The chi-square test column shows the results of the chi-square tests for independence, comparing sets 1 through 6 to the total distribution. * indicates p <.01, ** indicates p <.001.

Set	χ^2 (df=14)	p-value
1	13.68	.474
2	24.39	.041
3	11.49	.647
5	12.12	.597

Table 6: A subset of argument sets for $\mathcal{AF}_5$ compared against their total distribution.

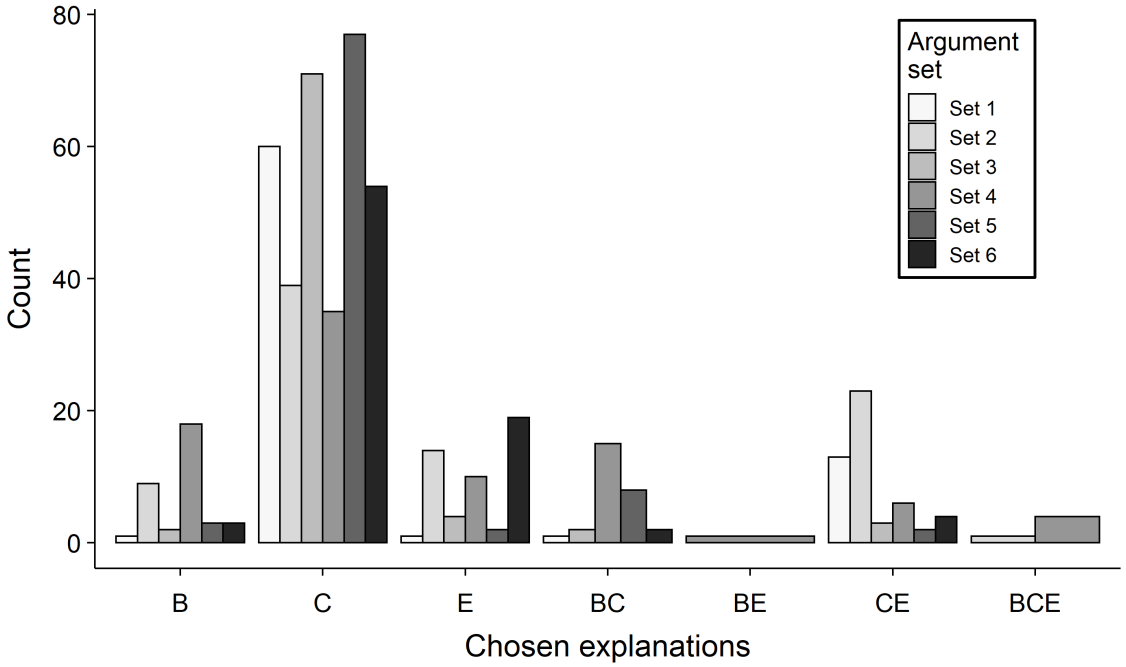


Figure 10: All explanations given for $\mathcal{AF}_4$ separated by argument set

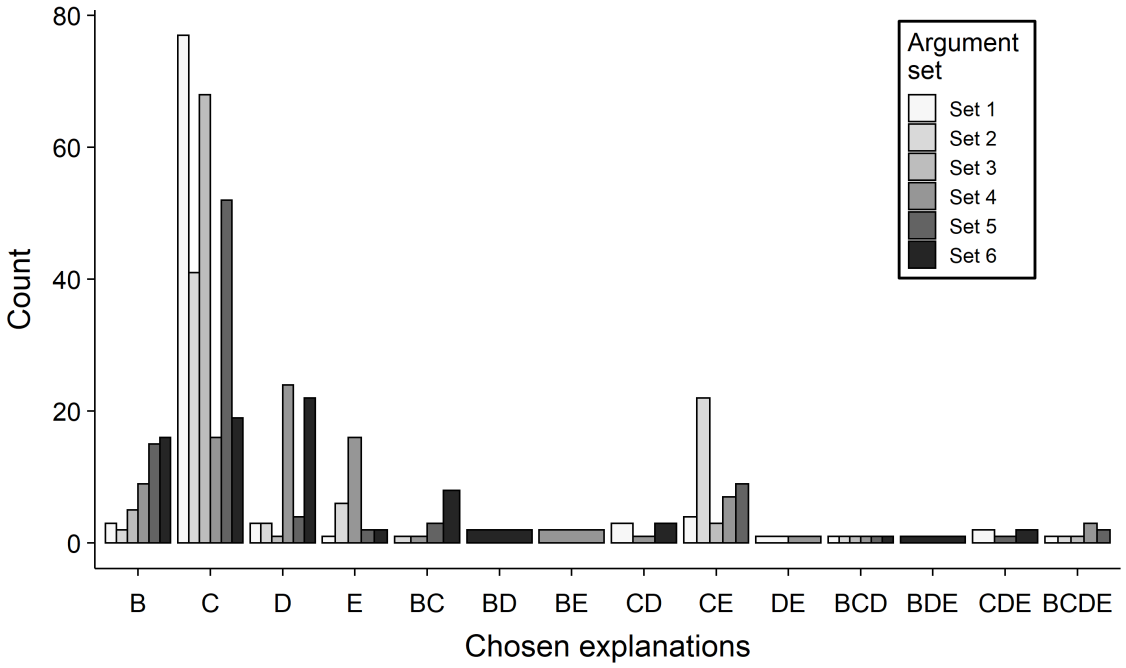


Figure 11: All explanations given for $\mathcal{AF}_5$ separated by argument sett

5.2 Sub-question 1: unrelated arguments

To test the first hypothesis, *"participants do not include unrelated arguments in their explanations"*, the explanations by participants were analysed in three steps. The explanations given were tested to see if they differed significantly from a random distribution over all explanation options. After this test, the data was grouped into three groups, explanations supporting the hypothesis, undermining the hypothesis, and other explanations. The proportion of each of these groups of the total explanations was calculated, after which the proportions were statistically compared. Based on the outcome of this test, individual explanation options were compared.

As can be seen in Figure 12, the explanations by participants do not appear to be randomly selected since there seems to be a pattern in the given explanations. A chi-square goodness-of-fit test was conducted to test if participants' explanations differed from chance. The test showed that the total distribution of the answers given by participants for $\mathcal{AF}_4$ (Table 4) differed from a uniform distribution ($\chi^2(6) = 1168.20$, $p < .001$).

Next, the explanations were grouped into three groups. The explantions including C and not E which support the hypothesis ($\{C\}$, $\{B, C\}$), explanations including E which undermine the hypothesis ($\{E\}$, $\{B, E\}$, $\{C, E\}$, $\{B, C, E\}$), and the remaining option $\{B\}$ which falls into neither category. The proportions of selected explanations that supported the hypotheses, undermined the hypothesis, and other explanations were calculated (Figure 13). Most of the provided explanations supported the hypothesis ($\hat{p} = .75$, $SE = 0.04$), a fifth of the explanations 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 if these proportions were significantly different from each other. More participants chose explanations aligning with the hypothesis than not aligned with the hypothesis ($\chi^2(1) = 249.19$, $p < .001$) and than explanations that fell into neither group ($\chi^2(1) = 437.74$, $p < .001$). The latter proportion was also lower than the proportion of explanations undermining the hypothesis ($\chi^2(1) = 50.78$, $p < .001$).

To further investigate the differences between the three groups of explanations, the proportions of all selected explanations were calculated and investigated to examine if there were differences within the three groups. The proportions of individual explanations were compared; the full results of all pair-wise comparisons of proportions can be found in Table 7. The explanations supporting the hypothesis were $\{C\}$ and $\{B, C\}$. As can be seen in Figure 14, $\{C\}$ was selected significantly more than all other explanations including $\{B, C\}$ ($\chi^2(1) = 486.15$, $p < .001$). Thus, the high proportion of hypothesis-supporting explanations is driven by C explanations. In the group of explanations including E, there was no difference in the number of times explanations $\{E\}$ and $\{C, E\}$ were chosen ($\chi^2(1) = 0.22$, $p = .643$). However, these two explanations were selected more often than all other explanations that included E.

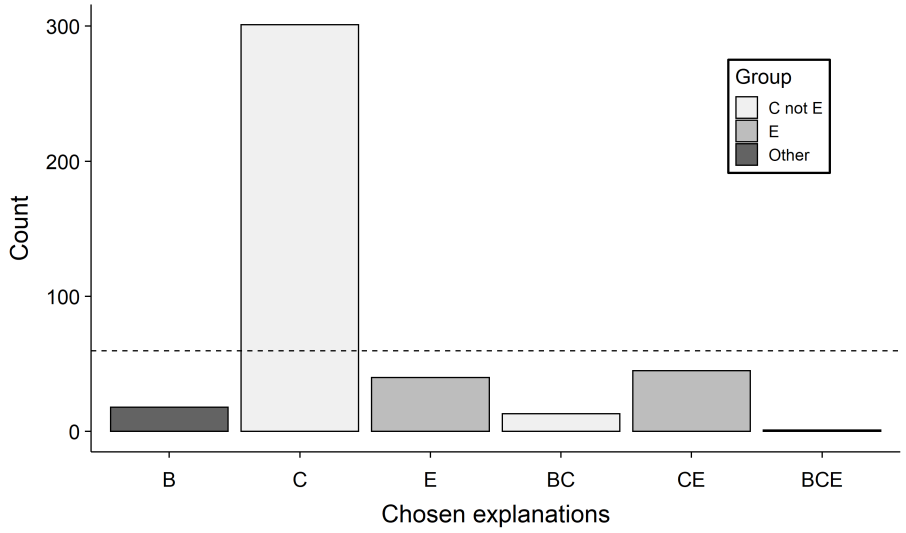


Figure 12: All explanations for $\mathcal{AF}_4$, excluding argument set 4. The dotted line indicated the expected counts if the explanations were uniformly distributed, which is $418/7 \approx 60$. BE was not selected so it is excluded from the plot.

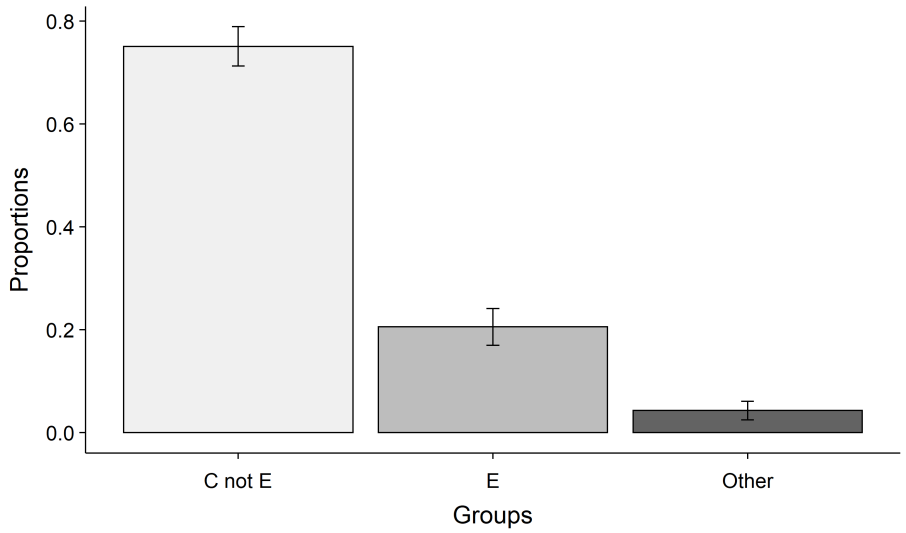


Figure 13: Explanations for $\mathcal{AF}_4$ grouped based on the hypothesis for sub-question 1, error bars indicate the standard error.

χ^2 (df=1)	B	C	E	BC	BE	CE
C	468.37**					
E	8.31*	396.48**				
BC	0.54	486.15**	13.83**			
BE	16.49**	535.22**	40.39**	11.29**		
CE	11.82**	381.46**	0.22	18.11**	46.06**	
BCE	13.86**	531.30**	37.47**	8.82*	0.00	43.10**

Table 7: Results of two proportion tests for $\mathcal{AF}_4$ ($df = 1$). * indicates p <.01, ** indicates p <.001.

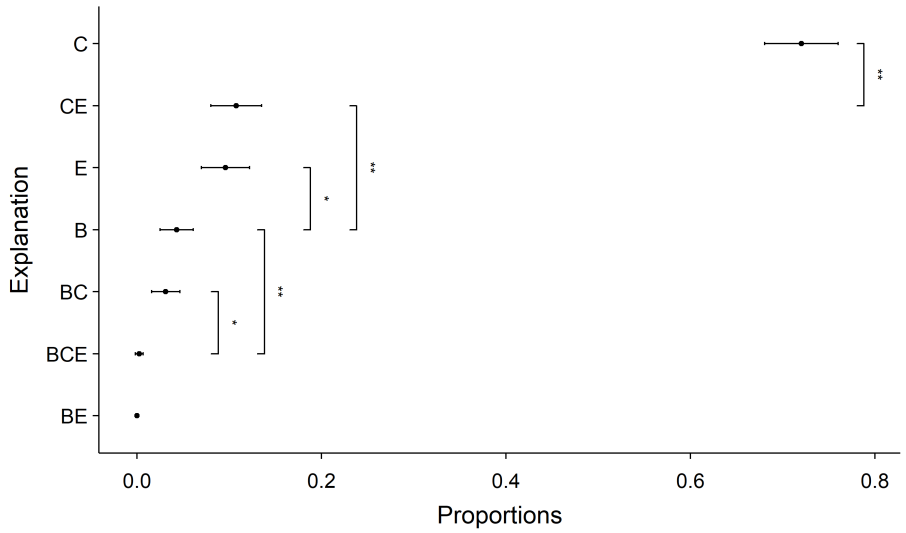


Figure 14: Proportions all explanations were selected in $\mathcal{AF}_4$, bars indicate the SE. Brackets have been used to indicate the closest explanation with which there is a significant difference; * indicates p <.01, ** indicates p <.001.

5.3 Sub-question 2: direct and indirect defenders

Based on the second sub-question, *"explanations were expected to include fewer indirect than direct defenders"*. This was tested using steps similar to those used for the first sub-question. The explanations were compared to chance. Data was grouped into three groups, explanations supporting the hypothesis, undermining the hypothesis, and other explanations. Then the proportions of these groups were compared. Based on the outcome of this test, individual explanations were compared.

Based on Figure 15, there appears to be a non-random pattern to explanations given by participants. A chi-square goodness-of-fit test was conducted to test if the participants' explanations differed from chance. The test showed that the total distribution of the explanations given by participants for $\mathcal{AF}_4$ (Table 4) differed from a uniform distribution ($\chi^2(14) = 2253.90$, p <.001).

The total distribution was grouped into three groups based on the hypothesis, and for each of these groups, their proportion of the total and standard error was calculated. Explanations including C and not E support the hypothesis ($\{C\}$ and $\{B, C\}$), explanations including E undermine the hypothesis ($\{E\}$, $\{B, E\}$, $\{C, E\}$ and $\{B, C, E\}$), all other explanations were considered to fall into neither group ($\{B\}$, $\{D\}$, $\{B, D\}$, $\{D, E\}$, $\{C, D\}$, $\{B, D, E\}$, $\{B, C, D\}$, $\{C, D, E\}$, and $\{B, C, D, E\}$). The proportions of selected explanations that fell into each of these three groups were calculated (Figure 16). Most of the provided explanations supported the hypothesis ($\hat{p} = .71$, $SE = 0.04$), a seventh of the explanations 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 aligning with the hypothesis than not aligned with the hypothesis ($\chi^2(1) = 228.33$, $p < .001$) and than explanations that fell into 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 ($\chi^2(1) = 0.30$, $p = .587$).

Since there was a difference between the three groups of explanations, the proportions of all selected explanations were calculated and compared to investigate differences within the three groups. The full results of all pair-wise comparisons of proportions can be found in 17, and all proportions, along with the next highest item from which they significantly differ, can be found in Figure 17. Two-proportion-tests were conducted to compare the proportions of individual explanations to gain insight into which options contributed most to the two relevant groups. Out of the two explanation options including C but not E, $\{C\}$ was selected significantly more than $\{B, C\}$ ($\chi^2(1) = 347.72$, $p < .001$). Out of the explanation options including E, the most frequently selected option $\{C, E\}$ was selected more than all other options that included E (Table 8). This means that $\{C\}$ and $\{C, E\}$ were the most common explanations and drove the majority of the differences between groups.

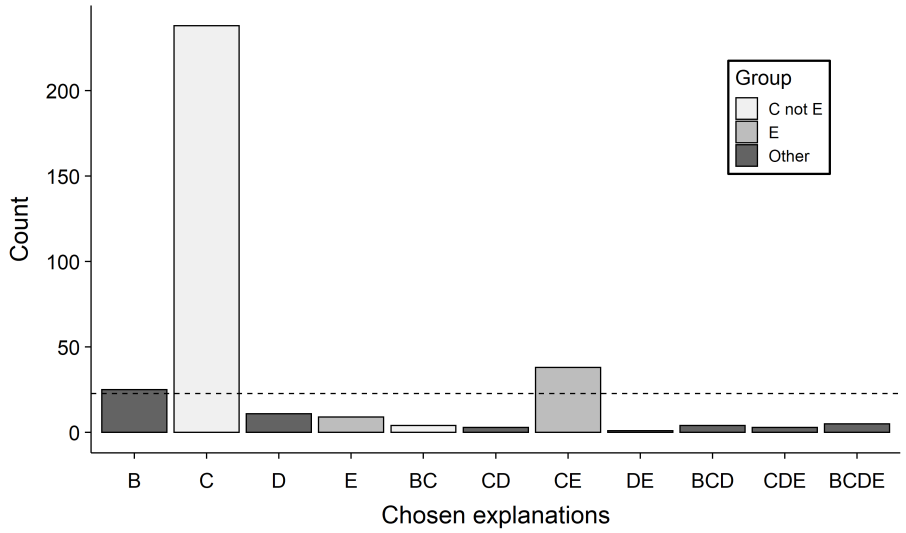


Figure 15: All explanations for $\mathcal{AF}_5$, excluding argument sets 4 and 6. The dotted line indicated the expected counts if the explanations were uniformly distributed, which is $341/15 \approx 23$. Explanations which were never selected were excluded from the figure.

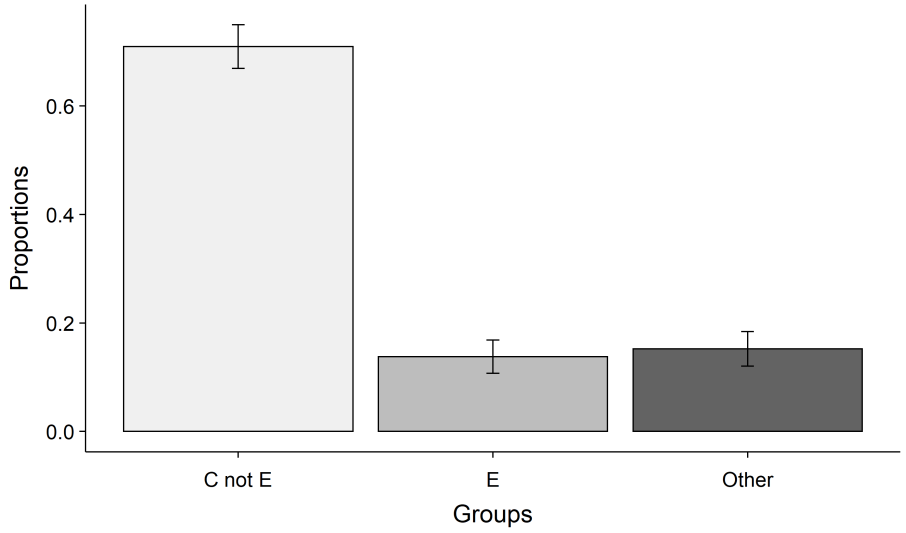


Figure 16: Explanations for $\mathcal{AF}_5$ grouped based on the hypothesis for sub-question 2, bars indicate the standard error.

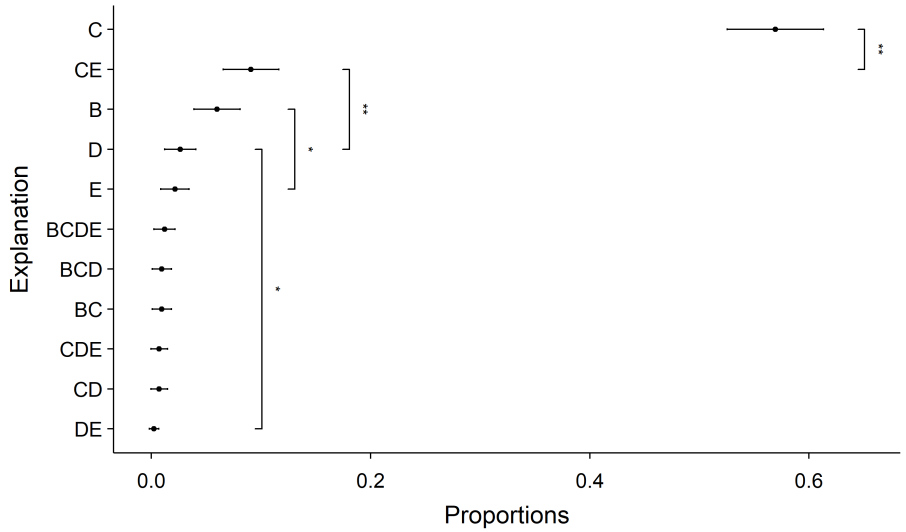


Figure 17: Proportions all explanations were selected in $\mathcal{AF}_5$, bars indicate the SE. Brackets have been used to indicate the closest explanation with which there is a significant difference; * indicates p <.01, ** indicates p <.001.

χ^2	B	C	D	E	BC	BD	BE	CD	CE	DE	BCD	BCE	BDE	CDE
C	278.16**													
D	4.96	323.08**												
E	6.97*	329.97**	0.05											
BC	14.41**	347.72**	2.45	1.26										
BD	23.92**	362.51**	9.24*	7.21*	2.26									
BE	23.92**	362.51**	9.24*	7.21*	2.26	0								
CD	16.42**	351.37**	3.57	2.12	0	1.34	1.34							
CE	2.52	241.02**	14.86**	17.92**	27.63**	38.15**	38.15**	23.00**						
DE	21.15**	358.76**	6.87*	4.97	0.81	0	0	0.25	35.25**					
BCD	14.41**	347.72**	2.45	1.26	0	2.26	2.26	0	27.63**	0.81				
BCE	23.92**	362.51**	9.24*	7.21*	2.26	0	0	1.34	38.15**	0	2.26			
BDE	23.92**	362.51**	9.24*	7.21*	2.26	0	0	1.34	38.15**	0	2.26	0		
CDE	16.42**	351.37**	3.57	2.12	0	1.34	1.34	0	30.00**	0.25	0	1.34	1.34	
BCDE	12.59**	344.10**	1.60	0.66	0	3.22	3.22	0.13	25.42**	1.51	0	3.22	3.22	0.13

Table 8: Results of two proportion tests for $\mathcal{AF}_5$ ($df = 1$). * indicates $p < .01$, ** indicates $p < .001$.

5.4 Differences based on demographics

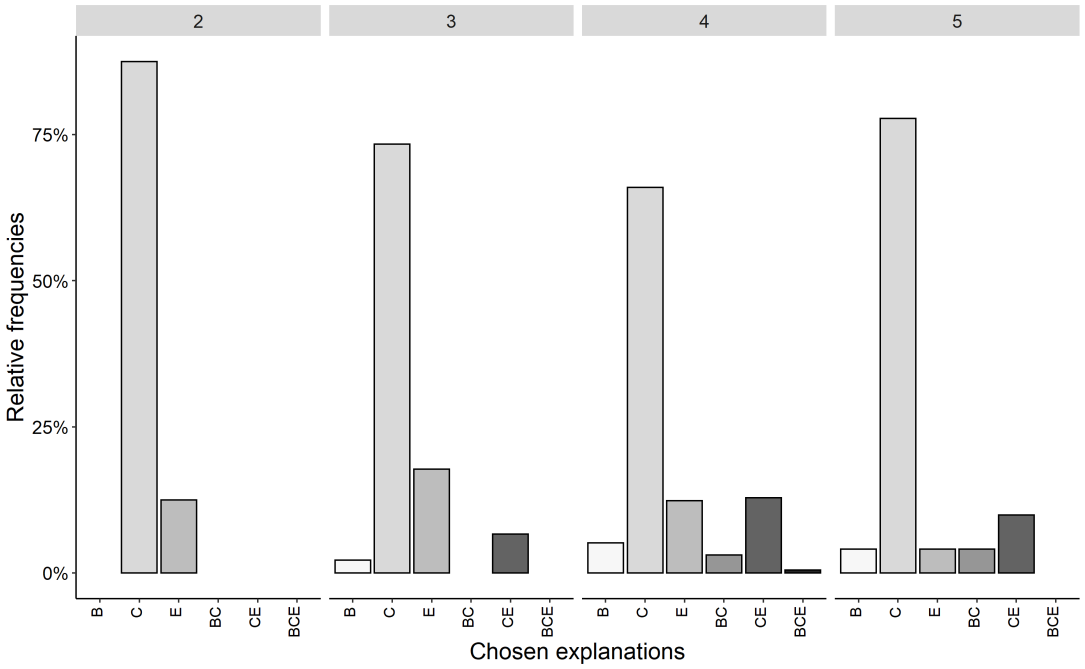
This study required careful reading of the different argument sets by participants, which they, based on comments left by participants, did not always find easy to do. Therefore, differences in provided explanations based on self-reported English proficiency were investigated. The rating participants gave themselves can be found in Table 9. As seen in Figure 18, there appears to be a preference for explanations consisting of only one argument for participants who were not very proficient in English (rating 2). However, it is important to note that only three participants (2.4 per cent) fell into this category. Thus no further conclusions can be drawn based on these results because of the small number of participants in this group.

The sample for this study contained participants with a wide range of familiarity with formal argumentation (Table 9). It is possible that participants with a better understanding of formal argumentation could grasp the AF underlying the argument sets better. This could cause differences in provided explanations based on familiarity with argumentation. An initial investigation of the given explanations separated by familiarity with argumentation shows a possible interaction effect in each AF. For $\mathcal{AF}_4$, both $\{B, C\}$ and $\{C, E\}$ appear to have been selected more by experts than by other groups. The proportions with which $\{B, C\}$ and $\{C, E\}$ were selected by experts were compared to the proportion they were selected by the other participants using two two-proportion tests. Experts did select $\{B, C\}$ ($\hat{p} = .17$) significantly more often than all other participants ($\hat{p} = .04$) ($\chi^2(1) = 13.87$, $p < .001$). No significant difference was found between the proportion of $\{C, E\}$ explanations by experts ($\hat{p} = .15$) and non-experts ($\hat{p} = .09$) ($\chi^2(1) = 1.27$, $p = .260$).

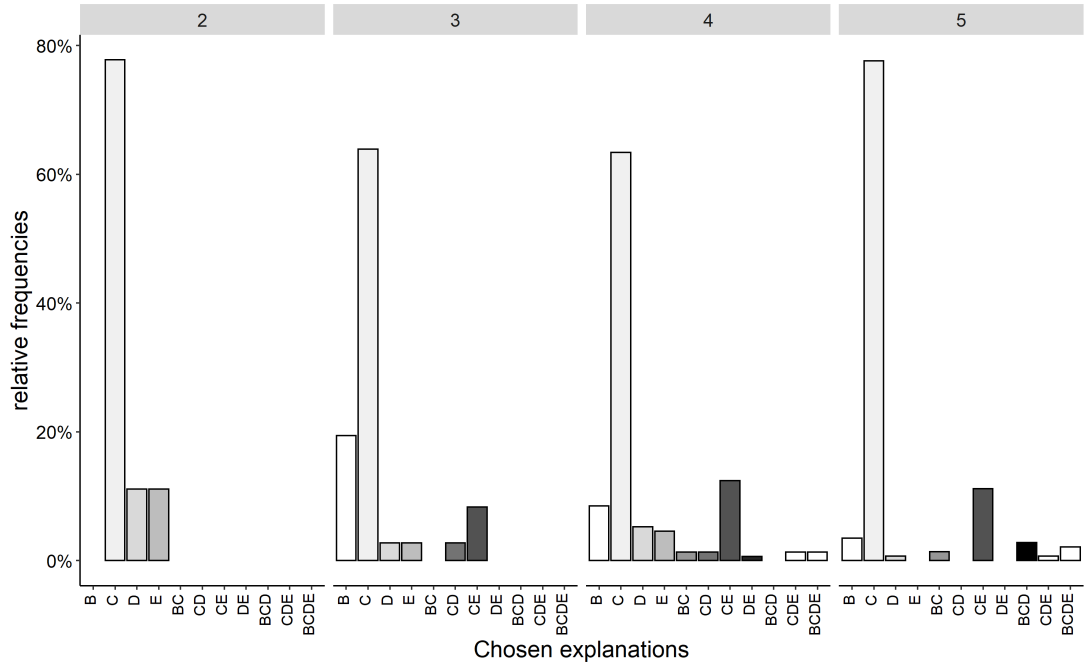
For $\mathcal{AF}_5$, there appears to be a difference in the frequency with which explanation $\{C, E\}$ was chosen for the different levels of familiarity. The largest difference in proportions between 'Expert' ($\hat{p} = .21$) and 'Quite familiar' ($\hat{p} = .04$) was found to be the only significant difference in proportions ($\chi^2(1) = 4.74$, $p = .029$). Experts selected $\{C, E\}$ more frequently than participants who were 'Quite familiar' with formal argumentation.

Familiarity with Argumenation			English proficiency (1-5)		
	Frequency	Percent (%)		Frequency	Percent (%)
Not heard of it	25	19.7	1 (not proficient)	0	0
Not familiar	38	29.9	2	3	2.4
Some familiarity	31	23.6	3	15	11.8
Quite familiar	19	15.0	4	57	44.9
Expert	14	11.0	5 (very proficient)	52	40.9

Table 9: Demographic information for all 127 participants.

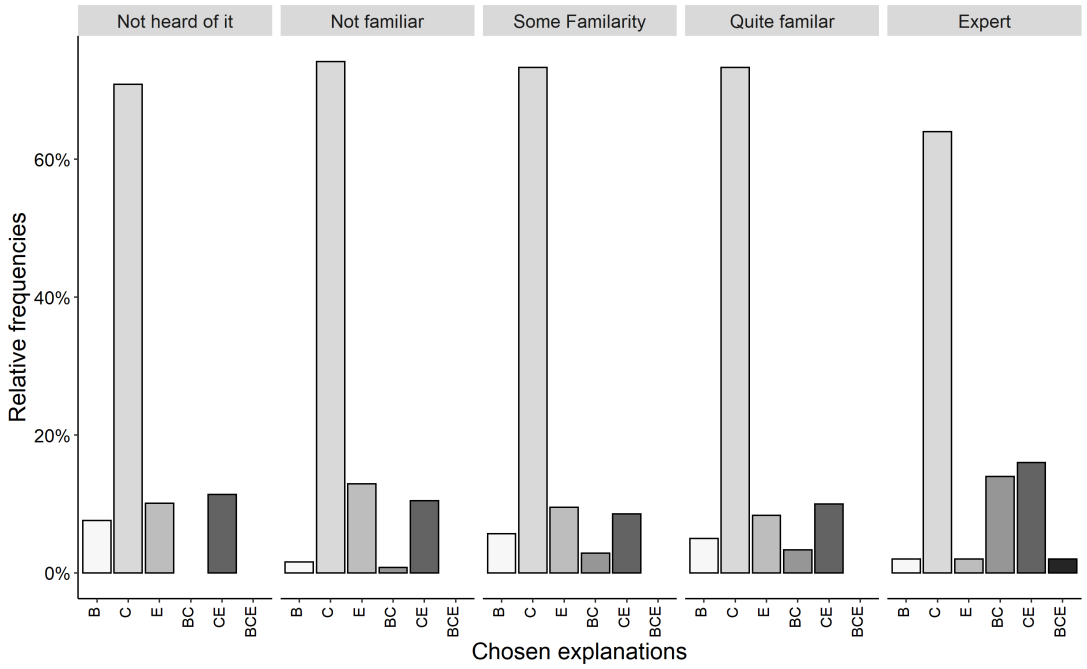


(a) $\mathcal{AF}_4$

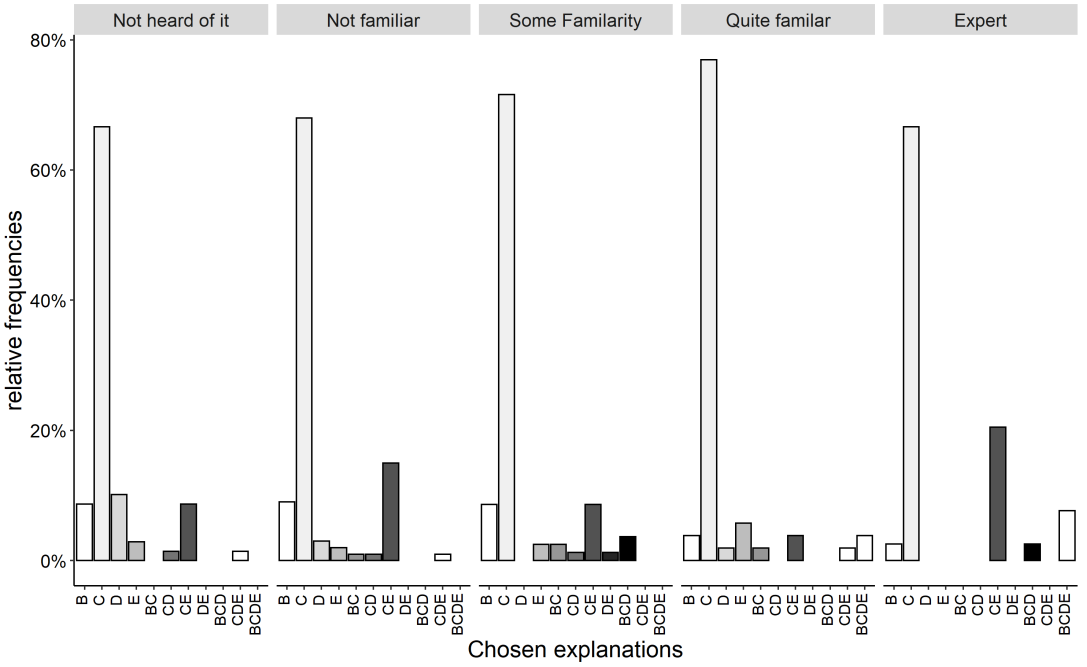


(b) $\mathcal{AF}_5$

Figure 18: Provided explanations separated by self-rated English proficiency where 1 indicated not proficient and 5 indicated very proficient.



(a) $\mathcal{AF}_4$



(b) $\mathcal{AF}_5$

Figure 19: Provided explanations separated by self-rated familiarity with computational argumentation.

6 Discussion

This study investigated the cognitive plausibility of explanations using formal argumentation to create an empirical basis for selecting between semantics in formal argumentation for XAI explanations. In this study, relevance, as used in formal argumentation, was compared to explanations provided by people. Based on the concept of relevance, people only include those elements in an explanation that are relevant to the topic being explained. The two sub-questions in this study each focused on different theories of relevance. The first sub-question, *"Do people include unrelated arguments in explanations?"*, aimed to test related admissible semantics as proposed by (Fan & Toni, 2014). It was hypothesised that participants would not include unrelated arguments since they selectively choose explanations, and unrelated arguments will not be selected since they do not include information relevant to the topic. The second sub-question, *"Do people prefer direct over indirect defenders?"*, focused on the relevance heuristic by Rosenfeld and Kraus (2016); according to this heuristic, people prefer arguments closer to the topic argument in explanation. Both sub-questions were addressed using an online observational study, in which participants were shown sets of arguments which instantiated two different AFs. These sets of arguments included a topic argument which participants were tasked to explain using any combination of the other arguments. In this section, the findings from this study will be discussed, interpreted, and placed in the wider context of formal argumentation and XAI. After this, the limitations of this study and some recommendations will be discussed, and a conclusion will be drawn.

6.1 Findings and interpretation

Evidence was found to support both hypotheses of this study. The first hypothesis, that *participants would not include argument E in explanations in $\mathcal{AF}_4$*, is supported by the finding that in most cases, they choose explanations that include C and not E. Both explanations $\{C\}$ and $\{B, C\}$ fit this description, but in most cases, participants chose $\{C\}$ as their explanation.

The hypothesis for the second sub-question, *participants prefer direct defenders of indirect defenders*, was supported by the finding that for $\mathcal{AF}_5$, participants chose explanations that included C (the direct defender) and not E (the indirect defender) significantly more than explanations including E.

In both AFs, $\{C, E\}$ was the second most commonly chosen explanation. However, this explanation was selected the same number of times as explanation $\{E\}$ in $\mathcal{AF}_4$ and B in $\mathcal{AF}_5$, neither of which are admissible explanations for the topic argument. Participants who answered E might not have understood the argument set or interpreted the task differently than intended; since there is no relation between E and the topic, argument E does not make sense as an explanation. Participants who chose B as an explanation might have chosen this option because it is the closest and directly related to the topic argument. However, argument B directly contradicts the topic argument and thus does not defend it. Therefore, based on the current data, it is difficult to say whether participants include E in explanation $\{C, E\}$ because they believe E should be included in the explanation or because they misunderstood the task or the argument set.

The primary research question of this study was *"Are explanations based on relevance in formal argumentation cognitively plausible?"*. The evidence for both sub-hypothesis supports a positive answer to this research question. Participants preferred not to include unrelated arguments in explanations, thus behaving as predicted by related admissible semantics. As predicted based on the relevance heuristic by Rosenfeld and Kraus (2016), participants preferred direct defenders over indirect defenders. These results show that these relevance-based measures are good descriptors of human behaviour, thus making them cognitively plausible. This study's findings align with the results found by Rahwan et al. (2010), showing that reinstatement is cognitively plausible. Participants could identify the direct defender of the topic argument even when arguments were shown in random order. Furthermore, participants primarily selected this direct defender to explain the topic, reinstating the topic argument. This shows that the findings by Rahwan et al. (2010) hold up even in a more complex context.

This study supports the use of related admissible semantics for generating explanations in XAI over semantics that do not select based on the relevance of arguments to a topic. This study also shows that people selectively choose explanations; most explanations included just one argument. Therefore, short selective explanations should perform well to mimic human explanations using AI systems.

It is important to note that in this study behaviour of participants was observed. Participants were tasked to show how they would explain the topic arguments based on little context. This does not mean

that these are the explanations that participants would prefer to receive from an AI system, nor that these are the explanations that best satisfy the goals of explanations in XAI. Participants might prefer to give shorter explanations for the sake of efficiency but would prefer longer explanations from an AI system that they have to trust and rely on for a task.

6.2 Limitations and future work

In this section, the limitations of this study, suggestions based on the challenges of this study, and possible future research directions will be discussed. The limitations of this study will be discussed in three main categories. These are the difficulties participants had with the task, the validity of this study, and the reliability of this study. Finally, some possible expansions to this study will be discussed.

6.2.1 Difficulty

The primary limitation of this study is that participants found the task very difficult. This was observed in all stages of the research, the pilot for the argument sets, testing the experiment, and comments left by participants in the experiment. This is a limitation because participants might have selected different options in a less difficult task since a more difficult task can increase the selectiveness of participants in choosing explanations (Cramer & Guillaume, 2019). It is possible that if participants had experienced the task as less difficult, they would have selected larger explanations, possibly selecting $\{C, E\}$ more often in $\mathcal{AF}_5$. The small number of participants who were the worst at English and likely found the task the most difficult did only select one option answer, providing evidence for this theory that difficulty increases selectiveness. Experts did select $\{C, E\}$ more than participants who were quite familiar with argumentation, however, the proportion with which experts selected this explanation did not differ from any other group. Therefore, this is not evidence that people who are more familiar with argumentation, who might have an easier time with the task, select larger explanations.

The randomization of argument order was likely a major contributor to the (perceived) difficulty of the task. When shown in order like in previous studies using similar argument sets (Rahwan et al., 2010; Cramer & Guillaume, 2019; Guillaume et al., 2022), each argument directly attacks the previous argument (unless the argument is unrelated), showing clear connections between all arguments. In an online study, it is vital to keep the participants' attention since the researcher has no control over the conditions in which the participants participate. Arguments were not shown in order for this study since it would have allowed participants to pick the same option for every explanation making the task easier over time, possibly losing their attention and focus. Another possible contributor to the perceived difficulty of the task in the study is the preconceived notion participants have of what a survey is. The study was conducted in an online survey format to make it easy to distribute the survey. Generally, surveys ask participants for their personal opinions and experiences and do not require much deliberation. Although resembling a survey, this study required much more thought and careful reading than participants might have expected. It is possible that participants found the study to be more complicated than they would have in another setting because they weren't prepared for a somewhat difficult task.

The task's difficulty in the study was primarily observed in the pilot and based on comments left on the study. Of course, not all participants reported finding the study difficult. Out of 127 participants, 14 commented on difficulties with specific argument sets or the entire study. The explanations given by participants were quite consistent, and the vast majority of explanations were the same. Therefore, it is possible that difficulty did not influence performance but instead, participants' confidence in their performance. To further investigate this, the (perceived) difficulty of the task would have to be reduced. One possible method of doing this is by providing participants with more context and a clearer role. Bezou Vrakatseli et al. (2021) included a generalized version of an AF before showing an exact instantiation. A similar approach could be used at the start of the experiment showing participants a large AF instantiated using natural language arguments. Participants get time to study and ask questions about this AF. Then a similar study to this one would follow using subgraphs of the AF as argument sets. Another approach to reducing the (perceived) difficulty of the task would be to include a group deliberation phase in the study design similar to Guillaume et al. (2022). In such a design, participants are first presented with the same choice of explanation themselves. After this, they discuss their answers with a small group and try to reach a consensus. Then participants get to provide another explanation for the same set on their own. Such a design reduces the cognitive load of the task by letting participants

help each other. If the difficulty of this task is primarily a lack of confidence in explanation, in such a study, participants would feel more confident in their performance but provide similar explanations to the current study. If the task's difficulty influenced the responses of participants in this study, they would respond differently in such a study.

The observation that participants found the study's task challenging is both a limitation and an interesting finding. An argument for using formal argumentation for explanations in XAI is that formal argumentation can represent how people talk, explain, and reason well (Atkinson et al., 2017; Vassiliades et al., 2021). However, in this experiment, people don't immediately intuitively understand the AF behind the arguments when shown an AF instantiated with natural language arguments. In the pilot used to verify the argument sets, participants correctly drew most of the attacks. However, not one of the five participants in the pilot figured out that they saw different instantiations of the same AF even after seeing more than 20 versions. This lack of identifying the underlying AF and the perceived difficulty with the task indicate that participants did not easily reason with the presented arguments. This might mean that argumentation does not align with human reasoning as well as previously speculated.

6.2.2 Validity

There are two threats to the validity of this study. First, this study uses a new study design. Therefore, there are no existing studies to compare the methods and results of this study. Comparing the result to another study would allow for more certainty in determining whether this study measured what it aimed to measure. This limitation primarily affects the unrelated arguments created for $\mathcal{AF}_4$. The inclusion of unrelated arguments in natural language argument sets is unprecedented. Therefore, all unrelated arguments had to be created for this study and could not be compared to existing material. The unrelated argument E was selected in set 2 more than in any other set, primarily in combination with argument C. It is possible that this unrelated argument was perceived as more related than other unrelated arguments leading to this difference in explanation behaviour between set 2 and other sets. Because only a small set of arguments was used in this study, no difference was made between different extents to which the unrelated arguments were unrelated. Where the line lies between an unrelated and related argument and whether participants' choices change based on the extent to which an argument is unrelated would have the investigated in a different study. Finding where this line lies would be valuable information for instantiating natural language versions of argumentation frameworks, including unrelated arguments.

The second threat to the validity of this study is that in the current study design, there is no way to determine whether participants understood the attack relations between arguments. They were not told about attacks between arguments besides that the sets consisted of 'arguments and counterarguments'. Participants in the pilot correctly indicated the majority of attacks between arguments, showing that they comprehended the structure between arguments. However, participants in the pilot were explicitly tasked to indicate attacks; participants in the main study were not asked about attacks. Thus, we cannot be entirely sure that participants understood attacks between arguments and used this information to choose explanations. However, all participants chose very similar explanations, which in the vast majority of cases included the direct defender to reinstate the topic argument. Therefore, even though it is unsure whether participants understood the AFs or attacks between arguments, they consistently chose the expected arguments.

6.2.3 Generalizibility

The generalizability of this study can be discussed in terms of four factors. Firstly, it is worth noting that because of the convenience sampling method used in the study, the sample is not fully representative of the general population. A relatively large number of participants, about half of the total sample, were at least somewhat familiar with formal argumentation. Most people in the general population are not familiar with formal argumentation. Participants in the sample also received a higher average level of education than the general population. This means that the findings of the study may not be entirely representative of the general population, particularly those less familiar with formal argumentation. The study included a limited number of participants with low self-rated English proficiency, making it difficult to draw any firm conclusions about this group. As a result, the generalizability of the study's findings to this population is also limited. Second, it is important to note that the study only considered acceptance explanations and did not investigate non-acceptance explanations. This means that the generalizability

of the study's findings to non-acceptance explanations is uncertain. Fan and Toni (2015b) define attack and argument explanations for non-acceptance, which only include related arguments. Based on the results of this study, these explanations are expected to be preferred over methods for non-acceptance explanations, which include unrelated arguments. This would need to be tested in a follow-up study. For this purpose, the argument sets, and instructions used in this study could be modified to apply to non-acceptance explanations. Third, the study focused on an extension-based method for explanation, which involves identifying sets of arguments that make a topic argument acceptable. This study did not investigate other methods of explanation, such as sub-graphs, labelling, and dialogue games. Therefore, the generalizability of the study's findings to these other methods of explanation is unclear. Finally, the AFs in this study were relatively small. This choice was made to not introduce too many distractions to answering the research question. Most participants chose the direct defender as their explanation of a topic argument. They could find this direct defender consistently despite the randomised argument order. It is possible that participants would also not have any trouble doing this for larger AFs.

Overall, the generalizability of this study is limited by the sample's familiarity with formal argumentation, the focus on acceptance explanations, and the exclusion of other methods of explanation. Further research is needed to explore these factors and their impact on the generalizability of the study's findings.

6.2.4 Future directions

Some interesting ideas for future study were identified during this study, including some gaps in the current body of research on the cognitive aspects of formal argumentation. Firstly, exploring prior information's effect on explanation preferences may be interesting. Many possible types of users can interact with XAI systems with varying degrees of prior knowledge. More knowledgeable users might prefer different types of explanations compared to less knowledgeable users (Miller, 2019). This could be explored by creating a novel setting for arguments of which no participant has prior knowledge and providing some participants with context on the setting. Alternatively, natural language argument sets could be created based on an existing system. Then explanation preferences could be compared between current users of the system and people who are unfamiliar with it.

Additionally, future research could examine the differences in explanation chosen for different goals, drawing on prior work such as the investigation of Tintarev and Masthoff (2015) into explanation goals for recommender systems. It would also be interesting to explore why participants chose specific explanations in the current study. This could, for example, reveal whether participants who chose $\{B\}$ were using counterfactual explanations. This could be achieved through a design that includes interviews with participants. More generally, a study design incorporating an interview element would be valuable because it can shed light on why participants make certain choices. Another potential direction for future research is to test other explanation generation methods in formal argumentation. The relevance of arguments is one method for choosing between extensions to use as an explanation. Other selection criteria that could be tested are compact and verbose explanations (Fan & Toni, 2014).

Important to note is that the arguments in this study were based on abstract argumentation frameworks and not based on any real system. More investigation is needed to bridge the gap between theoretical and practical applications of formal argumentation to generate explanations. A first set would be to repeat a similar study in which argument sets are based on an existing AI system or drawn for a structured argumentation system. Finally, future research could build on the current study by letting participants choose between two explanations for each topic to investigate how the effectiveness of different explanation methods varies depending on the options available. Alternatively, a two-step process could be used, where participants generate explanations, and others choose them, as has been done in prior research on recommender systems.

6.3 Conclusion

This study used a novel method to investigate the cognitive plausibility of explanations based on relevance in formal argumentation. Based on the results, relevance in argumentation seems to be cognitively plausible. Participants preferred small explanations consisting of direct defenders. However, further investigation is needed to determine whether the task's difficulty affects this study's results. Future work could build on the current work by expanding to non-acceptance and non-extension-based

explanations. And by investigating differences in explanation behaviour based on prior knowledge and goals of explanation.

References

Amgoud, L., & Cayrol, C. (1998). On the acceptability of arguments in preference-based argumentation. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (pp. 1–7). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks: The Official Journal of the International Neural Network Society*, *130*, 185–194. doi: 10.1016/j.neunet.2020.07.010

Angelov, P., Soares, E., Jiang, R., Arnold, N., & Atkinson, P. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*. doi: 10.1002/widm.1424

Antaki, C., & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, *22*(2), 181–194. doi: 10.1002/ejsp.2420220206

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi: 10.1016/j.inffus.2019.12.012

Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., ... Villata, S. (2017). Toward Artificial Argumentation. *AI Magazine*, *38*, 25–36.

Balog, K., & Radlinski, F. (2020). Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 329–338). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3397271.3401032

Baroni, P., Caminada, M., & Massimiliano, G. (2018). Abstract Argumentation Frameworks and Their Semantics. In *Handbook of Formal Argumentation* (pp. 159–236).

Baroni, P., & Giacomin, M. (2003). Solving Semantic Problems with Odd-Length Cycles in Argumentation. In T. D. Nielsen & N. L. Zhang (Eds.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (pp. 440–451). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-45062-7_36

Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., & Toni, F. (2014). Introduction to structured argumentation. *Argument & Computation*, *5*(1), 1–4. doi: 10.1080/19462166.2013.869764

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

Bondarenko, A., Dung, P., Kowalski, R., & Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, *93*(1-2), 63–101. doi: 10.1016/S0004-3702(97)00015-5

Borg, A., & Bex, F. (2021a). A Basic Framework for Explanations in Argumentation. *IEEE Intelligent Systems*, *36*(2), 25–35. doi: 10.1109/MIS.2021.3053102

Borg, A., & Bex, F. (2021b). Necessary and Sufficient Explanations for Argumentation-Based Conclusions. In J. Vejnarová & N. Wilson (Eds.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (Vol. 12897, pp. 45–58). Cham: Springer International Publishing. doi: 10.1007/978-3-030-86772-04

Briguez, C., Budán, M., Deagustini, C., Maguitman, A., Capobianco, M., & Simari, G. (2014). Argument-based mixed recommenders and their application to movie suggestion. *Expert Systems with Applications*, *41*, 6467–6482. doi: 10.1016/j.eswa.2014.03.046

Caminada, M. (2006). Semi-stable semantics. In *International Conference on Computational Models of Argument (COMMA 2006)* (pp. 121–130). IOS Press. doi: 10.1093/logcom/exr033

Caminada, M., & Wu, Y. (2011). On the Limitations of Abstract Argumentation. *Proc. 23rd Benelux Conf. on AI (BNAIC)*, 59–66.

Cayrol, C., & Lagasquie-Schiex, M. C. (2005). On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In L. Godo (Ed.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (pp. 378–389). Berlin, Heidelberg: Springer. doi: 10.1007/11518655_33

Cerutti, F., Tintarev, N., & Oren, N. (2014). Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation. *Frontiers in Artificial Intelligence and Applications*,

263, 212. doi: 10.3233/978-1-61499-419-0-207

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery*, *11*(1), e1391. doi: 10.1002/widm.1391

Cramer, M., & Guillaume, M. (2018a). Directionality of Attacks in Natural Language Argumentation. In C. Schon (Ed.), *Proceedings of the fourth Workshop on Bridging the Gap between Human and Automated Reasoning* (Vol. 2261, pp. 40–46). Stockholm, Schweden: CEUR.

Cramer, M., & Guillaume, M. (2018b). Empirical Cognitive Study on Abstract Argumentation Semantics. In *International Conference on Computational Models of Argumentation (COMMA 2018)* (pp. 413–424). IOS Press. doi: doi:10.3233/978-1-61499-906-5-413

Cramer, M., & Guillaume, M. (2019). Empirical Study on Human Evaluation of Complex Argumentation Frameworks. In F. Calimeri, N. Leone, & M. Manna (Eds.), *Logics in Artificial Intelligence* (pp. 102–115). Cham: Springer International Publishing. doi: 10.1007/978-3-030-19570-0_7

Cramer, M., & van der Torre, L. (2019). SCF2 – an Argumentation Semantics for Rational Human Judgments on Argument Acceptability. In C. Beierle, M. Ragni, F. Stolzenburg, & M. Thimm (Eds.), *Proceedings of the 8th Workshop on Dynamics of Knowledge and Belief (DKB-2019) and the 7th Workshop KI & Kognition (KIK-2019)* (Vol. 2445, pp. 24–35). Kassel, Germany: CEUR.

Čyras, K., Letsios, D., Misener, R., & Toni, F. (2019). Argumentation for Explainable Scheduling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 2752–2759. doi: 10.1609/aaai.v33i01.33012752

Čyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). *Argumentative XAI: A Survey* (No. arXiv:2105.11266). arXiv.

Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*(2), 321–357. doi: 10.1016/0004-3702(94)00041-X

Dung, P., Mancarella, P., & Toni, F. (2007). Computing ideal sceptical argumentation. *Artificial Intelligence*, *171*, 642–674. doi: 10.1016/j.artint.2007.05.003

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. doi: 10.1016/S1071-5819(03)00038-7

Fan, X., & Toni, F. (2014). On computing explanations in abstract argumentation. *Frontiers in Artificial Intelligence and Applications*, *263*, 1005–1006. doi: 10.3233/978-1-61499-419-0-1005

Fan, X., & Toni, F. (2015a). On Computing Explanations in Argumentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(1). doi: 10.1609/aaai.v29i1.9420

Fan, X., & Toni, F. (2015b). On Explanations for Non-Acceptable Arguments. In E. Black, S. Modgil, & N. Oren (Eds.), *Theory and Applications of Formal Argumentation* (Vol. 9524, pp. 112–127). Cham: Springer International Publishing. doi: 10.1007/978-3-319-28460-6_7

Gacutan, J., & Selvadurai, N. (2020). A statutory right to explanation for decisions generated using artificial intelligence. *International Journal of Law and Information Technology*, *28*, 193–216. doi: 10.1093/ijlit/eaaa016

Gaggl, S. A., & Woltran, S. (2013). The cf2 argumentation semantics revisited. *Journal of Logic and Computation*, *23*(5), 925–949. doi: 10.1093/logcom/exs011

García, A. J., Chesñevar, C. I., Rotstein, N. D., & Simari, G. R. (2013). Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Systems with Applications*, *40*(8), 3233–3247. doi: 10.1016/j.eswa.2012.12.036

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). doi: 10.1109/DSAA.2018.00018

Guillaume, M., Cramer, M., van der Torre, L., & Schiltz, C. (2022). Reasoning on conflicting information: An empirical study of Formal Argumentation. *PLOS ONE*, *17*(8), e0273225. doi: 10.1371/journal.pone.0273225

Hadoux, E., & Hunter, A. (2019). Comfort or safety? Gathering and using the concerns of a participant for better persuasion. *Argument & Computation*, *10*(2), 113–147. doi: 10.3233/AAC-191007

Hunter, A., & Polberg, S. (2017). Empirical Methods for Modelling Persuadees in Dialogical Argumentation. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 382–389). Boston, MA: IEEE. doi: 10.1109/ICTAI.2017.00066

Kennedy, W. G. (2009). Cognitive plausibility in cognitive modeling, artificial Intelligence, and social simulation. In *Proceedings of the International Conference on Cognitive Modeling (ICCM), Manchester, UK* (pp. 24–26).

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 1–16). Cham: Springer International Publishing. doi: 10.1007/978-3-030-57321-8_1

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. doi: 10.1017/S0140525X10000968

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. doi: 10.1016/j.artint.2018.07.007

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3351095.3372850

Odekerken, D., Bex, F. J., Villata, S., Harašta, J., & Křemen, P. (2020). Towards Transparent Human-in-the-Loop Classification of Fraudulent Web Shops. In *Legal Knowledge and Information Systems* (pp. 239–242). IOS Press.

Oren, N., van Deemter, K., & Vasconcelos, W. W. (2020). Argument-based plan explanation. *Knowledge Engineering Tools and Techniques for AI Planning*, 173–188. doi: 10.1007/978-3-030-38561-3_9

Polberg, S., & Hunter, A. (2018). Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *International Journal of Approximate Reasoning*, *93*, 487–543. doi: 10.1016/j.ijar.2017.11.009

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument & Computation*, *1*(2), 93–124. doi: 10.1080/19462160903564592

Prakken, H. (2018). Historical Overview of Formal Argumentation. In *Handbook of Formal Argumentation* (pp. 75–144).

Prakken, H., & de Winter, M. (2018). Abstraction in argumentation: Necessary but dangerous. In *International Conference on* (Vol. 305, pp. 85–96). doi: 10.3233/978-1-61499-906-5-85

Rahwan, I., Madakkatel, M. I., Bonnefon, J.-F., Awan, R. N., & Abdallah, S. (2010). Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement. *Cognitive Science*, *34*(8), 1483–1502. doi: 10.1111/j.1551-6709.2010.01123.x

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2939672.2939778

Rosenfeld, A., & Kraus, S. (2014). Argumentation theory in the field: An empirical study of fundamental notions. *CEUR Workshop Proceedings*, *1341*.

Rosenfeld, A., & Kraus, S. (2016). Providing Arguments in Discussions on the Basis of the Prediction of Human Argumentative Behavior. *ACM Transactions on Interactive Intelligent Systems*, *6*(4), 1–33. doi: 10.1145/2983925

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. doi: 10.1093/idpl/ipx022

Timmer, S. T., Meyer, J.-J. C., Prakken, H., Renooij, S., & Verheij, B. (2017). A two-phase method for extracting explanatory arguments from Bayesian networks. *International Journal of Approximate Reasoning*, *80*, 475–494. doi: 10.1016/j.ijar.2016.09.002

Tintarev, N., & Masthoff, J. (2007). A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop* (pp. 801–810). doi: 10.1109/ICDEW.2007.4401070

Tintarev, N., & Masthoff, J. (2015). Explaining Recommendations: Design and Evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 353–382). Boston, MA: Springer US. doi: 10.1007/978-1-4899-7637-6_10

Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review*, *36*, e5. doi: 10.1017/S0269888921000011

Verheij, B. (1996). Two Approaches to Dialectical Argumentation: Admissible Sets and Argumentation Stages. In *Proc. of the Eighth Dutch Conf. on Artificial Intelligence (NAIC'96)* (pp. 357–368).

Verma, T., Lingenfelder, C., & Klakow, D. (2020). Defining Explanation in an AI Context. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 314–322). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.29

Villata, S., Cabrio, E., Jraidi, I., Benlamine, S., Chaouachi, M., Frasson, C., & Gandon, F. (2017). Emotions and personality traits in argumentation: An empirical evaluation. *Argument & Computation*, *8*(1), 61–87. doi: 10.3233/AAC-170015

Winikoff, M., & Sardelić, J. (2021). Artificial Intelligence and the Right to Explanation as a Human Right. *IEEE Internet Computing*, *25*(02), 116–120. doi: 10.1109/MIC.2020.3045821

Yu, Z., Xu, K., & Liao, B. (2018). Structured Argumentation: Restricted Rebut vs. Unrestricted Rebut. *Studies in Logic*, *11*, 3–17.

Appendix A Pilot Study

A.1 Methods

The goals of the pilot were to test how people interpret the natural language argument sets for the two AFs discussed in the methods section, to test whether the sentences were easy to read even for non-native speakers and to test comprehension of arguments when presented in random order. For use in the study, the interpretation of the natural language argument sets by participants in the pilot should match the intended argumentation structure. 21 argument sets were tested for $\mathcal{AF}_4$ and 13 for $\mathcal{AF}_5$. Few sets were tested for the latter because some sets have already been tested in previous work (Cramer & Guillaume, 2018a) and because this task took participants longer.

The pilot was conducted using five participants between 20 and 30 years old. Two participants did not participate in the arguments for $\mathcal{AF}_5$. Participants were first given a short explanation of the task. Then participants were presented with all arguments in an argument set in random order. They were tasked to write down a letter for each argument (A, B, C, (D,) E) and draw an arrow from one argument to another if they believed there to be an attack from one to the other. The term 'attack' was not further explained to participants to test their naive interpretation of the AFs. Participants were allowed to draw any number of arrows (including none), and arguments were allowed to attack each other (bidirectional attack). If participants found sets of arguments particularly confusing or found what they perceived to be mistakes in arguments, they were encouraged to write this down or mention it to the researcher.

For $\mathcal{AF}_5$, a second slightly modified pilot study was performed with two participants. With the goal of reducing the number of bidirectional attacks indicated by participants. In this study, the participants completed the same task but were given additional clarification on the definitions of 'argument' and 'attack'. In this study, arguments (2) and (9), as numbered in Appendix D, were not included because these were found to be confusing by the participants.

The explanation given about the attacks was:

> Please indicate the 'attacks' between the following arguments. An argument is made up of a statement and a conclusion following that statement. An argument 'attacks' another argument if its conclusion directly contradicts the initial statement of another argument.
>
> An example:
> *A. The vase fell on the floor, so the vase broke.*
> *B. Daniel caught the vase, so it did not fall on the floor.*
>
> Here argument B attacks argument A because 'so it did not fall on the floor' directly contradicts 'the vase fell on the floor'. Argument A does not attack argument B since 'so the vase broke' does not directly contradict 'Daniel caught the vase'. Two arguments attack can also attack each other.

A.2 Findings

A.2.1 $\mathcal{AF}_4$

For $\mathcal{AF}_4$, the most frequently indicated attacks were the attacks from B to A (indicated 73.3 per cent of the time) and from C to B (77.1 per cent). These were also the two intended attacks based on the AF. The third and fourth most commonly indicated attacks were the counterparts of these attacks A to B (44.8 per cent) and from B to C (55.2 per cent). No other attack was indicated more than ten per cent of the time (Table A.1). This means that in the majority of the cases, participants agreed with the intended attacks, although participants often saw them as bidirectional. Although percentages for all other attacks were low, there were many argument sets in which participants did not see argument E as entirely unrelated. Out of the 21 sets for $\mathcal{AF}_4$, for nine, an attack to or from the 'unrelated' argument E was indicated by at least one participant. The other twelve argument sets had no attacks from or to E. One of these twelve arguments was confusing to several participants and was thus removed. One more argument set was dropped because of its similarity to other sets. Six of the remaining ten sets were used in the study and can be found in italics in Appendix B.

		To			
		A	B	C	E
From	A		44.8	1.0	3.8
	B	73.3		55.2	4.8
	C	2.9	77.1		2.9
	E	0.0	6.7	2.9	

Table A.1: Percentage of the time each possible attack was indicated by participants for $\mathcal{AF}_4$ with standard instructions.

Figure A.1 shows the attacks drawn by participants for $\mathcal{AF}_4$. By comparing the intended attacks listed at the top of the figure to the squared below, all attacks that were not intended can easily be spotted. Noteworthy are arguments (9) through (13) and (19) through (21) since for these argument sets, the attacks from A to B and from B to C seem to have been indicated less frequently than for other argument sets. The former are arguments from (Cramer & Guillaume, 2018b), and the latter are original arguments for this study.

A.2.2 $\mathcal{AF}_5$

For $\mathcal{AF}_5$, the most frequently indicated attacks were those between adjacent arguments, and both the intended and unintended directions were indicated frequently (Table A.2). In almost all cases (except from A to B), the intended attack was indicated more frequently than its counterpart. In 22 per cent of cases, an attack between B and E was indicated. In two argument sets, all participants indicated this attack. It is not strange for participants to indicate an attack between B and E since argument E is an indirect attacker of B. However, it was not the intended structure; thus, these two argument sets were removed. No participants indicated an attack between the other indirect attackers A and D.

Two major issues with the arguments for $\mathcal{AF}_5$ were found. One problem is that in many argument sets, participants did not indicate an attack from B to A but did indicate an attack from A to B. A second is the number of bidirectional attacks indicated by participants. Both of these observations suggest that the interpretation of the arguments by participants differs from the intended AF, and thus, these sets would not be usable for the study. Therefore, a second pilot was performed for $\mathcal{AF}_5$ with additional instructions on arguments and attack relations, as found in the methods section above.

The attacks indicated in the second pilot for $\mathcal{AF}_5$ can be found in Table A.3. In this study, the number of bidirectional attacks was greatly decreased compared to Table A.2. Thus, these instructions were helpful in aligning how people read the natural language arguments with the intended argumentation framework.

		To				
		A	B	C	D	E
From	A		47.0	0.0	0.0	0.0
	B	39.0		44.0	3.0	22.0
	C	3.0	58.0		53.0	11.0
	D	0.0	6.0	58.0		53.0
	E	0.0	8.0	11.0	61.0	

Table A.2: The percentage of the time each possible attack was indicated by participants for $\mathcal{AF}_5$ with standard instructions.

In Figure A.2 and Figure A.3 the attacks drawn by participants for $\mathcal{AF}_5$ can be seen. As can be seen, the number of indicated attacks was greatly reduced between the two versions of the pilot. Also noteworthy is that in Figure A.2, for argument sets (2) and (9) all participants indicated an attack from B to E. For the study 6 argument sets were chosen in which the interpretation by participants aligned most with the intended attacks. These argument sets can be found in Appendix D

		To				
		A	B	C	D	E
	A		10.0	0.0	0.0	0.0
	B	65.0		5.0	5.0	0.0
From	C	0.0	60.0		0.0	0.0
	D	0.0	0.0	40.0		0.0
	E	0.0	0.0	5.0	55.0	

Table A.3: The percentage of the time each possible attack was indicated by participants for $\mathcal{AF}_5$ with detailed instructions.

A.2.3 Number of attacks

A difference in the number of attacks between the two AFs could indicate that participants were more selective than the other. More attacks could be indicated by participants for $\mathcal{AF}_5$ than for $\mathcal{AF}_4$. For $\mathcal{AF}_4$ twelve different attacks were possible, for $\mathcal{AF}_5$ 25 attacks were possible. For $\mathcal{AF}_4$ there were two intended attacks (16.7 per cent of total options), and for $\mathcal{AF}_5$ four attacks were intended (16 per cent). For $\mathcal{AF}_4$ pilot participants indicated 22.9 per cent of attacks, for $\mathcal{AF}_5$ this was 23.9 per cent. So there was no major difference between the proportions of indicated attacks. In the second pilot study for $\mathcal{AF}_5$, the amount of indicated attacks was 12.3, approximately half compared to the version without additional instruction. This makes sense since almost all bidirectional attacks became unidirectional.

A.2.4 Participant feedback

Participants gave several useful notes to improve the wording of arguments. These primarily address some confusion arising from the shuffling of argument order. The argument sets taken from other studies were not originally intended to be shown in random order. Referencing words such as 'this' were sometimes confusing. Some argument sets where the timing of events was relevant were also complex to parse when presented in random order. In some argument sets important information, such as that Maxy is a pet snake, was only introduced in one of the 'first' arguments of a set, but when shuffled this information was introduced too late. Based on this feedback, minor wording fixes were done to improve the clarity of arguments. Participants indicated finding the task difficult; extra note was made of particular AFs that participants struggled with.

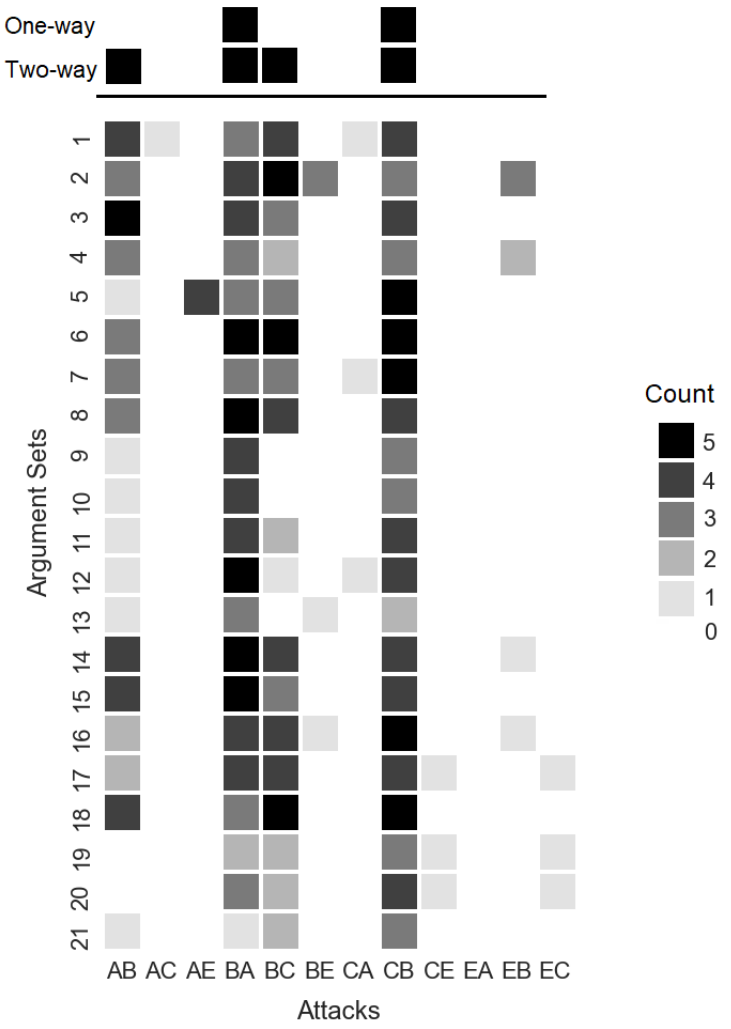


Figure A.1: All attacks drawn by participants for $\mathcal{AF}_4$. The first letter indicates the attacker, and the second indicates the attacked argument. AB stands for an attack from A to B.

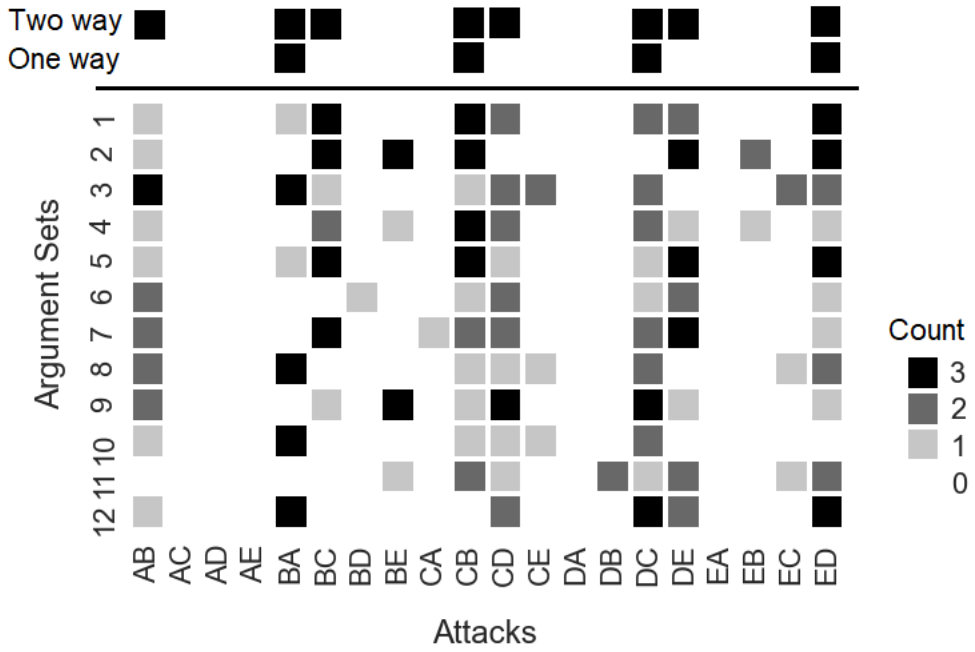


Figure A.2: All attacks drawn by participants for $\mathcal{AF}_5$ in the first pilot study. The first letter indicates the attacker, and the second indicates the attacked argument. AB stands for an attack from A to B.

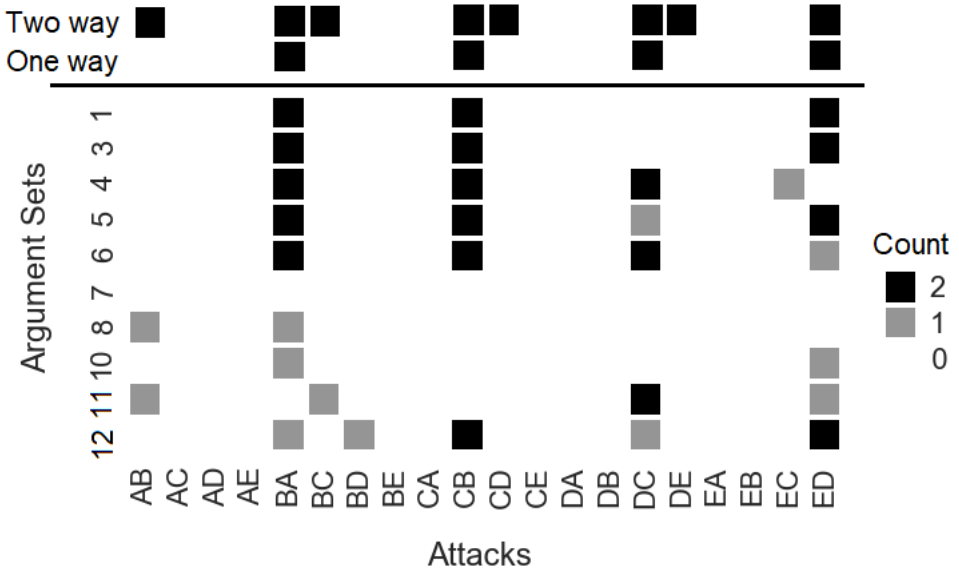


Figure A.3: All attacks drawn by participants for $\mathcal{AF}_5$ in the follow up pilot study. The first letter indicates the attacker, and the second indicates the attacked argument. AB stands for an attack from A to B.

Appendix B Arguments Used in Study

B.1 Arguments for $\mathcal{AF}_4$

	Source	Topic (A)	B	C	E
1	Rahwan et al., 2010	Stephen is not guilty. Therefore, Stephen is to be free from conviction.	Stephen was seen at the crime scene at the time of the crime. Therefore, Stephen is guilty.	Stephen was having dinner with his family at the time of the crime. Therefore, Stephen was not seen at the crime scene at the time of the crime.	Stephen is very tall. Therefore, Stephen likely doesn't leave small footprints.
2	Rahwan et al., 2010	Louis applied the brake and the brake was not faulty. Therefore, the car slowed down.	The brake fluid was empty. Therefore, the brake was faulty.	The car had just undergone maintenance service. Therefore, the brake fluid was not empty.	Louis is knowledgeable about car maintenance. Therefore, the brake fluid was not forgotten during maintenance.
3	Rahwan et al., 2010	The car did not slow down. Therefore, the car approached the signal at high speed.	Louis applied the brake. Therefore, the car slowed down.	Louis applied the accelerator instead. Therefore, Louis did not apply the brake.	Louis recently took the car to the garage. Therefore, there is nothing wrong with the car.
4	Cramer & Guillaume, 2018a	Bella says that there is a treasure buried to the north of the lake. So we should dig up the sand to the north of the lake.	Fiona says that Bella is not trustworthy, so we shouldn't follow Bella's directions.	Xavier says that Fiona is not trustworthy, so we shouldn't listen to Fiona's opinion on Bella.	Danny says that Vicky is not trustworthy, so we should not dig between the mountains like Vicky says.
5	Original	Peter paid for a product that never arrived, so he is a victim of fraud.	Peter did get a tracking code for the package, so the package will arrive.	The tracking code Peter received was fraudulent, so Peter never got a real tracking code for the package.	It's almost his partner's birthday, so Peter wants the package to arrive quickly.
6	Original	Johan got a passing grade on the exam and handed in the course assignment on time, so Johan will pass his course.	Johan handed in the course assignment past the deadline, so he did not hand in the assignments on time.	Johan got an extension for the assignment, so he did not hand it in past the deadline	Johan has missed two lectures, so he does not have 100 per cent attendance.

B.2 Arguments for $\mathcal{AF}_5$

	Source	Topic (A)	B	C	D	E
1	Rahwan et al., 2010	The battery of Alex's car is not working. Therefore, Alex's car will halt.	The battery of Alex's car has been changed today. Therefore, the battery of Alex's car is working.	The garage was closed today. Therefore, the battery of Alex's car has not been changed today.	Alex works at a garage. Therefore, Alex can change their car's battery even when the garage is closed.	Alex just got a new job. Therefore, Alex does not work at the garage anymore.
2	Rahwan et al., 2010	There is no electricity in the house. Therefore, all lights in the house are off.	There is a working portable generator in the house. Therefore, there is electricity in the house.	The fuel tank of the portable generator is empty. Therefore, the portable generator is not working.	The fuel tank of the generator was just refilled. Therefore, the tank is not empty.	The wrong fuel was put in the generator. Therefore, the tank was not properly refilled.
3	Rahwan et al., 2010	Cody does not fly. Therefore, Cody is unable to escape the zoo by flying.	Cody is a bird. Therefore, Cody flies.	Cody is a rabbit. Therefore, Cody is not a bird.	Cody does not have hair. Therefore, Cody is not a rabbit.	Cody was recently shaven. Therefore, Cody normally has hair.
4	Bezou Vrakatseli et al., 2021	The power is out, so Claire cannot charge her phone.	The heating is on, so the power is not out.	The thermostat is broken, so the heating is not on.	The thermostat was repaired, so the thermostat is not broken.	Not all parts for the repair were available, so the thermostat was not repaired.
5	Guillaume et al., 2022	John has built a fence around his garden, so his pet Tweety cannot escape.	Tweety is a bird and can fly, so the fence does not keep Tweety in.	Tweety has injured one of its wings by flying into a barbwire. So Tweety can no longer fly.	Tweety's injury happened a while ago, therefore Tweety is no longer injured.	Wings take a long time to heal compared to other body parts, so the injury did not happen that long ago.
6	Original	The lake is frozen so Dana can go ice skating tomorrow.	According to the weather broadcast the temperature is going to rise, so the lake won't be frozen tomorrow.	The temperature broadcast is unreliable, thus temperature might not rise.	This broadcast is part of a national news show, thus it is not unreliable.	The national news show and the weather broadcast have different funders and staff, thus they are not associated.

Appendix C Pilot Arguments $\mathcal{AF}_4$

	Source	Topic (A)	B	C	E
1	Rahwan et al., 2010	The battery of Alex's car is not working. Therefore, Alex's car will halt.	The battery of Alex's car has been changed at the garage. Therefore, the battery of Alex's car is working.	The garage was closed. Therefore, the battery of Alex's car has not been changed.	Alex got a new job. Therefore, Alex does not work at the garage anymore.
2	Rahwan et al., 2010	Louis applied the brake and the brake was not faulty. Therefore, the car slowed down.	The brake fluid was empty. Therefore, the brake was faulty.	The car had just undergone maintenance service. Therefore, the brake fluid was not empty.	Louis is knowledgeable about car maintenance. Therefore, the brake fluid was not forgotten during maintenance.
3	Rahwan et al., 2010	Mary does not limit her phone usage. Therefore, Mary has a large phone bill.	Mary has a speech disorder. Therefore, Mary limits her phone usage.	Mary is a singer. Therefore, Mary does not have a speech disorder.	Mary keeps work and private life separate. Therefore, Mary has two phones.
4	Rahwan et al., 2010	John has no way to know Leila's password. Therefore, Leila's e-mails are secured from John.	Leila's secret question is very easy to answer. Therefore, John has a way to know Leila's password.	Leila purposely gave a wrong answer to her secret question. Therefore, Leila's secret question is not very easy to answer.	John is not very smart. Therefore, John can't guess Leila's password.
5	Rahwan et al., 2010	Mike's laptop does not have anti-virus software installed. Therefore, Mike's laptop is vulnerable to computer viruses.	Nowadays anti-virus software is always available by default on purchase. Therefore, Mike's laptop has anti-virus software.	Some laptops are very cheap and have minimal software. Therefore, anti-virus software is not always available by default.	Mike is knowledgeable about cyber security. Therefore, Mike won't click on any harmful links.
6	Rahwan et al., 2010	Cody does not fly. Therefore, Cody is unable to escape by flying.	Cody is a bird. Therefore, Cody flies and can escape by flying.	Cody is a rabbit. Therefore, Cody is not a bird.	Cody swims a lot. Therefore, Cody has strong legs.
7	Rahwan et al., 2010	The car did not slow down. Therefore, the car approached the signal at high speed.	Louis applied the brake. Therefore, the car slowed down.	Louis applied the accelerator instead. Therefore, Louis did not apply the brake.	Louis recently took the car to the garage. Therefore, there is nothing wrong with the car.
8	Rahwan et al., 2010	Stephen is not guilty. Therefore, Stephen is to be free from conviction.	Stephen was seen at the crime scene at the time of the crime. Therefore, Stephen is guilty.	Stephen was having dinner with his family at the time of the crime. Therefore, Stephen was not seen at the crime scene at the time of the crime.	Stephen is very tall. Therefore, Stephen likely doesn't leave small footprints.
9	Cramer & Guillaume, 2018a	Irina says that there is a treasure buried near the northern tip of the island. So we should dig up the sand near the northern tip of the island.	Hans says that Irina is not trustworthy so we should not trust Irena.	Peter says that Hans is not trustworthy, so we should not trust what Hans says.	Jenny says that Dennis is not trustworthy, so we should not trust Dennis.
10	Cramer & Guillaume, 2018a	Peter says that there is a treasure buried near the southern tip of the island. So we should dig up the sand near the southern tip of the island.	Livia says that Peter is not trustworthy, and that there is a treasure buried near the eastern tip of the island. So we should not trust what Peter says, and we should dig up the sand near the eastern tip of the island.	Jenny says that Livia is not trustworthy and that there is a treasure buried on the peak of the mountain. So we should not trust what Olivia says, and we should dig up the sand on the peak of the mountain.	Kaylee says that there is a treasure buried near the northern tip of the island. So we should dig up sand at the northern tip of the island.

11	Cramer & Guillaume, 2018a	Vincent says that there is a treasure buried next to the ruins. So we should dig up the sand next to the ruins.	Ursula says that Vincent is not trustworthy, so we shouldn't dig up the sand next to the ruins.	Tom says that Ursula is not trustworthy, so we should trust what Vincent says.	Ron says that Sarah is not trustworthy and that we shouldn't listen to Sarah.
12	Cramer & Guillaume, 2018a	Bella says that there is a treasure buried to the north of the lake. So we should dig up the sand to the north of the lake.	Fiona says that Bella is not trustworthy, so we shouldn't follow Bella's directions.	Xavier says that Fiona is not trustworthy, so we shouldn't listen to Fiona's opinion on Bella.	Danny says that Vicky is not trustworthy, so we should not dig between the mountains like Vicky says.
13	Cramer & Guillaume, 2018a	Anna says that there is a treasure buried to the west of the lake. So we should dig up the sand to the west of the lake.	Zoe says that Anna is not trustworthy, and that there is a treasure buried between the two highest mountains. So we should not trust what Anna says, and we should dig up the sand between the two highest mountains.	Xavier says that Zoe is not trustworthy and that a treasure is buried next to the temple. So we should not trust what Zoe says, and we should dig up the sand next to the temple.	Yanis says that Walter is not trustworthy and that a treasure is buried to the east of the lake. So we should not trust what Walter says, and we should dig up the sand to the east of the lake.
14	Bezou Vrakatseli et al., 2021	The power is out, so Claire cannot charge her phone.	The TV is playing, so the power is not out.	The TV is broken, so the TV is not playing.	The phone is broken, so the phone is not ringing.
15	Guillaume et al., 2022	Patient A is infected with the Norovirus. The 2002 Encyclopedia of Medicine states that antibiotics can treat the Norovirus. So patient A can be healed with antibiotics.	A peer-reviewed research article by Gold et al. from 2007 has established that antibiotics cannot treat the Norovirus. Therefore no patient infected with the Norovirus can be healed with antibiotics.	A study that the Medical School of Harvard University has published in 2013 corrects mistakes made in the study by Gold et al. and concludes that only cyclic antibiotics can treat Norovirus.	A peer-reviewed research article by Blume et al. from 2008 has established that antibiotics can treat bacterial infections.
16	Guillaume et al., 2022	Marry put Maxy in a large cage, so Maxy cannot escape.	Maxy is a tiny snake, so Maxy can escape through the holes of its cage.	The cage is specifically intended for tiny snakes; so all holes in the cage are too small for Maxy to go through.	Maxy has been growing a lot, so Maxy has just shed their skin.
17	Guillaume et al., 2022	Mark locked Quicky in his room, so Quicky cannot escape.	Quicky is a rodent, so Quicky can escape through the gap below the door.	Quicky is a large guinea pig, so Quicky cannot go through the gap below the door.	Quicky loves to eat carrots, so Quicky eats enough vegetables
18	Original	Peter paid for a product that never arrived, so he is a victim of fraud.	Peter did get a tracking code for the package, so the package will arrive.	The tracking code Peter received was fraudulent, so Peter never got a real tracking code for the package.	It's almost his partner's birthday, so Peter wants the package to arrive quickly.
19	Original	The lake is frozen so Dana can go ice skating tomorrow.	According to the weather broadcast the temperature is going to rise, so the lake won't be frozen tomorrow.	The temperature broadcast is unreliable, thus temperature might not rise.	Dana just bought new skates, so the skates will be sharp.
20	Original	Daisy says there is a treasure buried underneath the crooked tree.	Timmy says that Daisy is not trustworthy, so the treasure is not where she indicated.	Harold says that Daisy is very kind and helpful, therefore she is trustworthy.	Harold and Timmy are on the same sports team, so they see each other frequently.
21	Original	Johan got a passing grade on the exam and handed in the course assignment on time, so Johan will pass his course.	Johan handed in the course assignment past the deadline, so he did not hand in the assignments on time.	Johan got an extension for the assignment, so he did not hand it in past the deadline	Johan has missed two lectures, so he does not have 100 per cent attendance.

Appendix D Pilot Arguments $\mathcal{AF}_5$

	Source	Topic (A)	B	C	D	E
1	Rahwan et al., 2010	The battery of Alex's car is not working. Therefore, Alex's car will halt.	The battery of Alex's car has been changed today. Therefore, the battery of Alex's car is working.	The garage was closed today. Therefore, the battery of Alex's car has not been changed today.	Alex works at a garage. Therefore, Alex can change their car's battery even when the garage is closed.	Alex just got a new job. Therefore, Alex does not work at the garage anymore.
2	Rahwan et al., 2010	Mike's laptop does not have anti-virus software installed. Therefore, Mike's laptop is vulnerable to computer viruses.	Nowadays anti-virus software is available by default on laptops. Therefore, Mike's laptop has anti-virus software.	Some laptops are very cheap and have minimal software. Therefore, anti-virus software is not always available by default.	Mike works in computer science. Therefore, Mike's laptop does not have minimal software.	Mike recently opened his restaurant. Therefore, Mike does not work in computer science.
3	Rahwan et al., 2010	There is no electricity in the house. Therefore, all lights in the house are off.	There is a working portable generator in the house. Therefore, there is electricity in the house.	The fuel tank of the portable generator is empty. Therefore, the portable generator is not working.	The fuel tank of the generator was just refilled. Therefore, the tank is not empty.	The wrong fuel was put in the generator. Therefore, the tank was not properly refilled.
4	Rahwan et al., 2010	Cody does not fly. Therefore, Cody is unable to escape the zoo by flying.	Cody is a bird. Therefore, Cody flies.	Cody is a rabbit. Therefore, Cody is not a bird.	Cody does not have hair. Therefore, Cody is not a rabbit.	Cody was recently shaven. Therefore, Cody normally has hair.
5	Bezou Vrakatseli et al., 2021	The power is out, so Claire cannot charge her phone.	The heating is on, so the power is not out.	The thermostat is broken, so the heating is not on.	The thermostat was repaired, so the thermostat is not broken.	Not all parts for the repair were available, so the thermostat was not repaired.
6	Cramer & Guillaume, 2018a	Davey says that there is a treasure buried next to the ruins. So we should dig next to the ruins.	Ursula says that Davey is not trustworthy and that a treasure is buried next to the old monument. So we should dig next to the old monument instead of the ruins.	Janet says that Ursula is not trustworthy and that a treasure is buried south of the forest. So we should dig south of the forest instead of next to the monument.	Henry says that Janet is not trustworthy and that a treasure is buried north of the forest. So we should dig up the sand north instead of south of the forest.	Ron says that Henry is not trustworthy and that a treasure is buried next to the harbour. So we should dig up the sand next to the harbour instead of north of the forest.
7	Cramer & Guillaume, 2018a	Ivan says that a treasure buried is to the north of the village. So we should dig to the north of the village.	Hannah says that Ivan is not trustworthy, so we should dig up the sand to the south of the village instead.	Emma says that Hannah is not trustworthy, so we should not follow Hannah's directions.	Dorothy says that Emma is not trustworthy and that a treasure is buried to the west of the high mountain. So we should not trust what Emma says, and we should dig to the west of the high mountain instead.	Charlie says that Dorothy is not trustworthy and that a treasure is buried to the south of the high mountain. So we should dig up the sand to the south of the high mountain instead.
8	Guillaume et al., 2022	Specimen A consists only of amylase. The 2003 Encyclopedia of Biochemistry states that amylase is not an enzyme. So specimen A does not contain any enzymes.	The International Institute for the Advancement of Biochemistry published new research results in 2006 that show that amylase is an enzyme. Therefore any specimen consisting of amylase contains an enzyme.	According to an article in the New York Times, the International Institute for the Advancement of Biochemistry does not exist, and the publications published under this name are a hoax. Therefore these publications cannot be trusted.	The article in the New York Times about the International Institute for the Advancement of Biochemistry was retracted in a later edition since it contained erroneous information. Therefore, the article cannot be trusted.	The retraction of the article about the International Institute for the Advancement of Biochemistry was not published by the New York Times, but instead by a website imitating the New York Times. Therefore, information published by this source should not be trusted.
9	Guillaume et al., 2022	Dana put their pet snake Maxy in a large cage, so Maxy cannot escape.	Maxy is a tiny snake, so Maxy can escape through the holes of its cage.	The cage is specifically intended for tiny snakes; so all holes in the cage are too small for small snake Maxy to go through.	The cage is advertised to also hold large spiders, so the cage is not specifically intended for snakes.	To hold large spiders the roof of the cage needs to be upgraded, so the cage cannot hold large spiders.

10	Guillaume et al., 2022	John has built a fence around his garden, so his pet Tweety cannot escape.	Tweety is a bird and can fly, so the fence does not keep Tweety in.	Tweety has injured one of its wings by flying into a barbwire. So Tweety can no longer fly.	Tweety's injury happened a while ago, therefore Tweety is no longer injured.	Wings take a long time to heal compared to other body parts, so the injury did not happen that long ago.
11	Original	The lake is frozen so Dana can go ice skating tomorrow.	According to the weather broadcast the temperature is going to rise, so the lake won't be frozen tomorrow.	The temperature broadcast is unreliable, thus temperature might not rise.	This broadcast is part of a national news show, thus it is not unreliable.	The national news show and the weather broadcast have different funders and staff, thus they are not associated.
12	Original	Daisy says there is a treasure buried underneath the crooked tree.	Timmy says that Daisy is not trustworthy, so the treasure is not where she indicated.	Harold says that Daisy is very kind and helpful, therefore she is trustworthy.	Harold and Daisy are married, thus Harold is not a reliable source for Daisy.	Harold and Daisy recently got divorced, so they are no longer married.

Appendix E Research Participant Information Sheet

Welcome to the Study 'Explanation based on Formal Argumentation'

By Roos Scheffers, Graduate School of Natural Sciences, Utrecht University. Supervisor: Floris Bex

1. Introduction
My name is Roos, and I am a graduate student at Utrecht University. I'm conducting a study on explanations in the field of explainable artificial intelligence as part of my master's thesis under the supervision of Floris Bex. You are invited to take part in this study. Before you decide it is important to understand why the research is being done and what it will involve.

2. About the study
Models created using artificial intelligence can be very complex. The field of explainable artificial intelligence aims to explain these very complex models and to create functional models with reduced complexity. To be able to explain models to people it is important to understand what explanations people prefer and use. This study will collect such data and compare it to theories of explanation based on formal argumentation.

3. Your rights
Participation is voluntary. Your data for this study will only be collected if you consent to this. If you decide not to participate, you do not have to take any further action. You do not need to sign anything. Nor are you required to explain why you do not want to participate. If you decide to participate, you can always change your mind and stop participating at any time, including during the study. You will even be able to withdraw your consent after you have participated. However, if you choose to do so, we will not be required to undo the processing of your data that has taken place up until that time. The personal data we have obtained from you up until the time when you withdraw your consent will be erased (where personal data is any data that can be linked to you, so this excludes any already anonymized data).

4. Approval of this study
This study has been allowed to proceed by the Research Institute of Information and Computing Sciences based on an Ethics and Privacy Quick Scan. If you have a complaint about the way this study is carried out, please send an email to: ics-ethics@uu.nl. If you have any complaints or questions about the processing of personal data, please send an email to the Faculty of Sciences Privacy Officer: privacy-beta@uu.nl. The Privacy Officer will also be able to assist you in exercising the rights you have under the GDPR. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.

If you have any questions or concerns about this research please contact Roos Scheffers at r.j.scheffers@uu.nl or my supervisor or my supervisor Floris Bex at f.j.bex@uu.nl.

Please read the statements below and tick the final box to confirm you have read and understood the statements and upon doing so agree to participate in the project.

- I confirm that I am 18 years of age or over.
- I confirm that the research project "Explanation based on Formal argumentation" has been explained to me.
- I consent to the material I contribute being used to generate insights for the research project.
- I understand that my participation in this research is voluntary and that I may withdraw from the study at any time without providing a reason, and that if I withdraw any personal data already collected from me will be erased.
- I consent to allow the fully anonymized data to be used in future publications and other scholarly means of disseminating the findings from the research project.
- I understand that the data acquired will be securely stored by researchers, but that appropriately

anonymized data may in future be made available to others for research purposes.

- I understand that the University may publish appropriately anonymized data in appropriate data repositories for verification purposes and to make it accessible to researchers and other research users.

Appendix F Study Instructions

Please read the instructions for the next part of the experiment carefully.

In this study, you will be presented with 8 scenarios, each of which consists of several arguments and counterarguments. The topic argument, which will be at the top of the page, can always be taken to be true. The other arguments will be below it. Your task will be to explain why the conclusion of the topic argument is the case. Choose the options that you feel explain the topic argument's conclusion best, by clicking the box next to the arguments. You can select any number of arguments to explain the topic argument.
For this study, I'm interested in how you personally would choose to explain the topic. There are no right or wrong answers; you can explain each scenario however you feel is best. Please select the options that best correspond to how you would explain the topic argument.
There is no time limit on this task so feel free to take as long as you need to read all arguments carefully. Every question will include an info button that you can click to see a shortened version of this explanation. Feel free to take a look at the first question and go back to the instructions if you want to read them again. You can go back to these instructions at any time.

Appendix G Experiment Layout

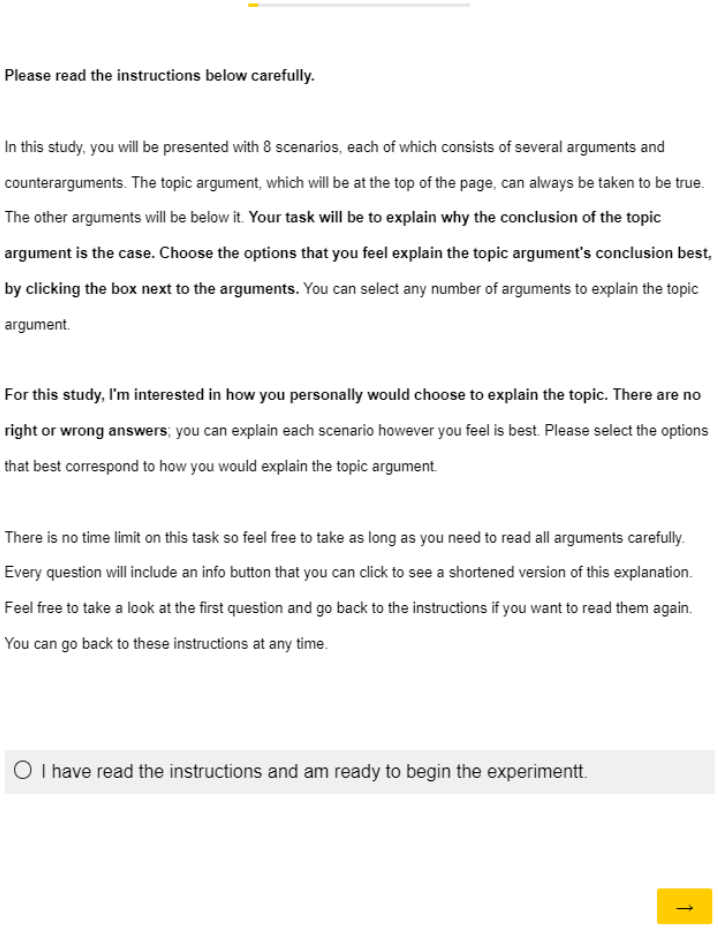
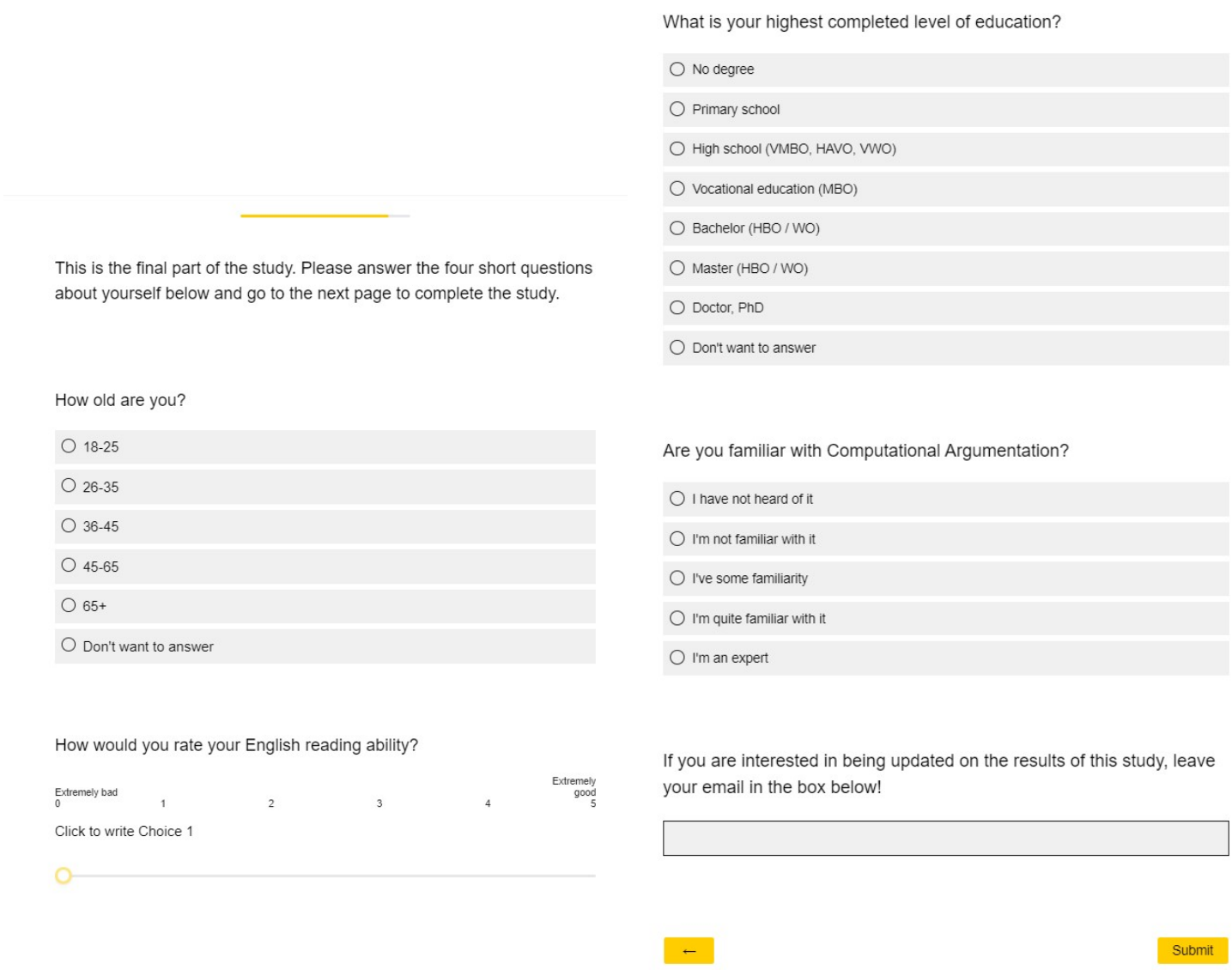


Figure G.1: Instructions of the study



(a) (b)

Figure G.2: Demographics questions included at the end of the survey, (a) is presented above (b).

Appendix H Ethics and Privacy Quick Scan

Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- Yes

Recruitment

P2. Does your project involve participants younger than 18 years of age?

- No

P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent?

- No

P4. Is your project likely to involve participants engaging in illegal activities?

- No

P5. Does your project involve patients?

- No

P6. Does your project involve participants belonging to a vulnerable group, other than those listed above?

- No

P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients?

- No

Informed consent

PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorized representatives? (See suggestions for information sheets and consent forms on the website.)

- Yes

PC2. Will you tell participants that their participation is voluntary?

- Yes

PC3. Will you obtain explicit consent for participation?

- Yes

PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings?

- Not applicable

PC5. Will you tell participants that they may withdraw from the research at any time and for any reason?

- Yes

PC6. Will you give potential participants time to consider participation?
- Yes

PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)?
- Yes

PC8. Does your project involve concealment or deliberate misleading of participants?
- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?
- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?
- No

H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?
- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)
- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)
- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?
- No

H6. Does your project give rise to a realistic risk of harm to the researchers?
- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?
- No

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?
- No

H9. Is there a realistic risk of other types of negative externalities?
- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information. Please see the guidance on the ICS Ethics and Privacy website on what happens on submission.

Z0. Which is your main department?

- Information and Computing Science

Z1. Your full name:

Rosalie Johanna Scheffers

Z2. Your email address:

r.j.scheffers@students.uu.nl

Z3. In what context will you conduct this research?

- As a student for my master thesis, supervised by::
  Floris Bex

Z5. Master programme for which you are doing the thesis

- Artificial Intelligence

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):

f.j.bex@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):

supportoffice.ai.ics@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:

Relevant Explanations in Formal Argumentation, an Empirical Study

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):

In this thesis I will investigate whether theoretically relevant explanations in based on formal argumentation theory are comparable to explanations selected by people in an experiment. With as goal to determine whether these theories in formal argumentation based on relevance align with how humans explain things.
In the experiment, which will take the form of a survey, participants will be shown a topic and various options which could be used to explain the topic. Participants will be tasked to choose those option that they believe form an appropriate explanation for the topic. The options chosen by participants will be compared to the options predicted based on theory.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

- Not applicable

Scoring

- Privacy: 0
- Ethics: 0