Exploring Contrastive Explanations in Formal Argumentation

Modelling contrastiveness based on literature in the social sciences

Sophie Glade (6224709)

Supervisor: Dr. AnneMarie Borg
Second Supervisor: Prof. dr. Floris J. Bex

Master Thesis
Artificial Intelligence

Faculty of Science
Utrecht University
The Netherlands
May 2023

Abstract

With the growing usage of artificial intelligence (AI) in daily life, explainable systems become more important. Explainable AI (XAI), which is a set of tools and frameworks to help you understand and interpret predictions made by AI, has risen in popularity due to this. Formal argumentation is a suitable tool that can be used to model contrastive explanations for (X)AI systems. In this thesis, the concept of contrastiveness will be modelled in various formal argumentation settings. The definition for contrastiveness is based on definitions found in literature from the field of the social sciences and humanities. First, an extensive literature research is conducted to find suitable definitions for contrastiveness, then those concepts are modelled in Dung's abstract argumentation frameworks, preference-based argumentation frameworks and finally in the structured argumentation setting of ASPIC$^+$. These definitions are then evaluated at the hand of examples based on real-life situations. This work successfully models various notions of contrastiveness in various formal argumentation settings and paves the way for the further study and application of contrastiveness in (argumentative) XAI.

Contents

1 Introduction

Artificial Intelligence (AI) has become a prominent field in recent years [Chu17]. It is being deployed across many different sectors, such as the medical field [MPR+19] and in legal analysis [Ash17]. Along with the growth of AI itself, there has been a rapid increase in popularity of Explainable Artificial Intelligence (shortened to XAI) as well. XAI is the field in which explanation techniques are researched and developed, and it paves the way to explain the decisions made by elaborate and difficult to understand processes in AI ([Mit97], [VL21b]). The goal of XAI is to make decisions of AI systems understandable for every user. The user can always wonder why a certain decision is made, but an explanation helps them to understand the reasons behind the decision. The AI system must be able to give explanations that provide insight into the underlying decision making processes, so that users can understand, trust, and validate the system. Not to mention that this way experts can verify that the system works as intended [DBH18]. There have been many different methods to go about explaining decisions made by AI systems ([BBM+15], [MP14a], [Rai20] and [Rud19], to name a few), to achieve more transparent systems.

This thesis will focus on formal argumentation as the main method for achieving explanations for AI systems. Formal argumentation can manage debatable information and can draw conclusions using formalised arguments. Argumentation also has its own applications ([OBBT22], [ABG+17]), and has been applied successfully to generate explanations for learning-based approaches. It is important to know how we can generate good explanations from formal argumentation, to then be able to implement it on a wider scale.

Much research has been done in the social sciences on what kind of explanations humans tend to prefer ([Har65], [MK93], [Hit99], [BW02], [Ove11], among others). Miller reviewed much of this research and found that explanations are often contrastive [Mil19]. That is, people do not ask why event P happened, but rather why event P happened instead of some other event Q: when people only ask *"Why P?"*, they often mean *"Why P rather than Q?"*. There is a contrast present between that which did happen and that which did not. Since it has been shown that people tend to ask questions in a contrastive manner, it is important for XAI systems to be capable of giving explanations to questions of that format.

1.1 Background

1.1.1 XAI

In recent years, the field of Artificial Intelligence (AI) has grown substantially. From being a term coined first by John McCarthy in 1956 [Moo06], AI has become a widespread phenomenon. Most people come across AI in their daily lives, in applications such as a digital assistant in the form of Apple's Siri, Amazon's Alexa or the Google Assistant, or in algorithms that provide you with content it thinks would suit you best [BFG11], such as the widely used app TikTok that curates a front page for you based on what you view on the app [Zha21]. However, as the use of AI becomes more widespread, it also becomes important that when such a system makes a decision, it is a decision that we as humans can understand and trust. While getting recommended something on your TikTok *'For You Page'* that you are not particularly interested in is not especially harmful, as the stakes are quite low in that case, we would not want these systems to produce faulty output when the stakes are much higher, such as in a court of law or with a medical diagnosis. Before being able to use AI in fields such as law enforcement and medicine, we need to make sure these AI systems make fair and trustworthy choices.

This brings us to the field of Explainable Artificial Intelligence (XAI). XAI is AI in which explanation techniques are researched and developed so that humans can understand the decisions made by the AI system [VL21b]. Due to the growth of relevancy of AI, the topic of XAI has also seen an increase in popularity in the last decade, with more publications and conferences on the topic [AB18]. XAI needs AI to follow the three principles of transparency, interpretability and explainability [DBH18]. An approach is transparent if the processes can be described and motivated by the designer of that approach [RBDG20]. An approach is considered interpretable if it can present some of the properties in terms which are understandable to a human [RBDG20]. Unlike the first two, there is not one widely accepted definition for explainability, but in this research it will be defined as the following: an approach is explainable if it is able to provide a way to improve the understanding of the user [CH19].

AI systems need to be transparent, interpretable and explainable to not just the experts who made the systems (which is then called a white-box [VL21a]), but also to lay-people; non-experts who may not understand all intricacies of certain models. This is necessary if AI is to be used by everyone in their daily life.

There are a multitude of reasons as to why XAI is important and it will only grow to become more important as the field of AI itself keeps growing [XUD+19]. First of all, XAI is important to people who *use* AI systems. When an AI system recommends a decision, the user needs to *understand* the underlying process to be able to trust that decision, and handle accordingly. For example, a medical doctor needs to know the reason why an algorithm prescribed a certain cause to inputted symptoms of a patient, before making additional decisions. XAI is also important to people who are *affected* by the decisions of AI systems. Consider the previous example: the patient would need to understand why a specific diagnosis was given to them. Otherwise, the patient might not feel safe or adequately treated. And finally, XAI is also important for the purpose of *improving* AI systems. If the underlying processes are clear and explainable, problems such as bugs and biases can be fixed or altogether prevented. In the same medical example, you would not want someone's race or gender to unknowingly and wrongly influence the diagnosis due to a bias in the system. One could not even start improving such a system when it is unknown how it works to begin with. All of these reasons point to why XAI is essential if we want users to gain trust in AI and have them continue using AI systems, which is a necessity given the increasing prevalence of AI in various fields.

1.1.2 Formal Argumentation

There is a need to provide a way to improve the understanding of AI of the user, and the ability to generate explanations is key to achieve this. A good explanation needs to reveal the underlying reasoning processes in terms interpretable to a human [ZMLC18]. A way to achieve this, is through argumentation. Argumentation is the study of how reasonable decisions or conclusions can be reached by constructing for and against arguments and evaluating these arguments accordingly ([ZMLC18]) and formal argumentation is a manner of logical inference that is based on constructing and evaluating arguments, each of which provides reasons for a particular claim. Logical models of argument are a tool to formalise reasoning and have contributed to a better understanding of reasoning problems in AI [CML00].

Explanations are often found to be argumentative [AL92]. As a result, argumentation is a promising field to do more research in, and it is becoming a more common tool in XAI applications. Argumentation has been succesfully applied in many different fields; current research addresses a range of applications in law, medicine, e-government and debating ([ABG+17], [CML00], [OBBT22]). With explanations often being argumentative, it is important to know how we can generate good explanations from argumentation for it to become more widely usable. Progress in the field of argumentation is expected to contribute to significant advances in the understanding and modeling of various aspects of human intelligence [ABG+17]. That is why in this thesis, the decision has been made to focus on argumentation as a way to provide explanations for decisions made by an AI system.

AI systems should be able to handle incomplete and inconsistent information, in a way similar to how humans would handle it. The way humans generally do this, is by argumentation; either internally (evaluating arguments) or externally (entering a discussion) [ABG+17]. Arguing for something can be deemed as an explanation, as argumentation is capable of transparently explaining the procedure and the results of reasoning [FT15]. From a theoretical point of view, explanation and argumentation are two notions that are relatively close and difficult to distinguish [Tho73]. Both explanation and argumentation are used equally in problem solving, and thus systems should also be equipped with explanation and argumentation capabilities in order to convince their users of the validity of their choices [MIBD02].

Formal argumentation can roughly be split into two approaches: abstract argumentation and structured argumentation. Mostly popularised in [Dun95] in 1995, abstract argumentation makes use of argumentation frameworks, which are graphs consisting of a set of abstract arguments (the nodes), and binary attack relations between those arguments. Ever since, many have expanded on this basic notion of abstract argumentation (see [Pra18] for an overview). One such expansion is known as Preference-Based Argumentation, which accounts for preferences within the argumentation framework [AC02]. Structured argumentation, on the other hand, makes use of a formal language for representing knowledge, and it specifies how arguments can be constructed from that knowledge. Arguments are generally constructed so that the premises and claims of the argument are made explicit, and the relationship between premises and claims is formally defined (for example, by using logical entailment) [BGH$^+$14]. One such framework is known as ASPIC$^+$ [MP14b]. ASPIC$^+$ provides a formal and computational foundation for studying and analyzing structured argumentation in a systematic manner, and has been widely used in the field of artificial intelligence, particularly in areas such as automated reasoning, multi-agent systems, and argument-based decision-making.

From the above it follows that formal argumentation is a suitable tool for explanation in XAI and it makes sense to apply it, as argumentation theory focuses on the human perspective of explanation. As a result, the use of argumentation within the field of AI and XAI has become more widespread (see [CRA$^+$21] for an overview). Progress in the field of formal argumentation can aid in advancing the understanding of various aspects of human intelligence, and thus the significance of formal argumentation in the field of XAI should not be overlooked.

1.1.3 Contrastiveness

As has been stated in Section 1, most explanations are contrastive. This section will elaborate more on the topic of contrastiveness and its importance.

The most common reason for people to ask questions is when something unexpected has happened [Gar80]. The type of questions asked in such unexpected situations are found to always be *contrastive* questions ([HM19], [BW02]): when wanting an explanation for an occurrence, people often ask the question *"Why did P happen?"*. By asking questions of this form, what people truly want is an explanation to the question of *"Why did P happen instead of Q?"*. In this case, P is the *fact*; that which did happen, and Q is the counterfactual case, the so-called *foil*; that which has not happened or is not the case [Lip90]. Even if the foil is left implicit, this construction is a *contrastive question*, and to answer it a *contrastive explanation* is needed, in response to the counterfactual case [Mil19]. In other words, there is a contrast present between the fact and the foil. To then answer this question, it is important to explain as many of the differences between fact and foil as possible [Lip90].

Something to note here is that there exists a difference between counterfactual and contrastive explanations. For this research, I will make a distinction between the two, and will follow the same definition given in [SACP21]: contrastive explanations point to the difference between the actual and a hypothetical decision, while counterfactual explanations specify necessary minimal changes in the input so that a contrastive output is obtained. Not all researchers keep that distinction in mind, and some use the terms contrastive and counterfactual interchangeably. It is important to consider this when literature is discussed in the following sections, as some previous research might have slightly different opinions on what is contrastive and what is counterfactual. While counterfactuals are not the topic of this research, some papers that discuss counterfactuals are still included, as their definitions of counterfactuals actually match my definition of contrastiveness.

This work will also occasionally refer to *causal explanations*. Causality on its own is too broad of a topic to discuss in detail, considering the scope and focus of this research. For the sake of clarity, however, some elaboration is useful. When speaking of causality and causal explanations, this thesis will adhere to the following definition: one could imagine there to be a *history* of events, which can be conceived as a list of particular happenings, and one (or multiple) of those events is responsible, or is the cause, for a specific outcome - in contrast to a counterfactual outcome. Fact and foil both have their own history of events leading up to how the current state came into existence. For the sake of this thesis, this is enough to know about causality with regards to explanations.

Returning to the main topic of contrastiveness, the definition of a contrastive *question* is not something that is widely disputed. Generally it is agreed that a contrastive question constitutes of the form of *"Why P rather than Q?"*. However, the definition of a contrastive *explanation* is not as readily available. There are some ideas of contrastive explanations that multiple researchers do agree on. For example, a good contrastive explanation should consist of stating the contrast between the fact and the foil ([Lip90], [Hil90], [Hit99], [BW02], [CPB17], [HM19] and [MK93], among others). Yet, that is where most similarities end. How does one find the contrast between the fact and the foil? Many researchers have their own way of defining contrastive explanations, and put the emphasis of what constitutes a good contrastive explanations on different aspects. There is not one single consensus of what is a contrastive explanation.

Taking into account all of the aforementioned, studying contrastiveness in a setting of formal argumentation, which can be used for XAI, is vital. Especially as it can be shown that not much research has been done on the topic as of today ([SACP21], [BB22b]). That is why, in this thesis, I will research contrastiveness within the field of formal argumentation based on findings in the social sciences and humanities.

1.2 Research Questions

This brings us to the topic of this research, and the main research question that this research will focus on:

> *Based on different definitions given in existing social sciences literature, how can contrastiveness be modelled in formal argumentation?*

To answer this research question and to differentiate from previous research, I will research not only how contrastiveness is defined across existing literature in the field of the social sciences and humanities, but also how these different definitions can be modelled in various argumentation settings - both abstract and structured argumentation. These definitions will then also be evaluated at the hand of several examples based on real-world scenarios. The goal is to create definitions that return explanations that are accessible.

The main research question will be divided into different subquestions, which will help give structure to this research. In this thesis, I will answer the following suquestions:

1. *What are various definitions given in the fields of humanities and the social sciences for contrastiveness and what similarities can be concluded?*
2. *How can these definitions of contrastiveness be modelled in an abstract argumentation setting?*
3. *How can these definitions be modelled and evaluated in preference-based argumentation?*
4. *How can these definitions be modelled and evaluated in ASPIC$^+$?*

1.3 Outline

To answer the main research question and its subquestions, this thesis is structured as follows: first, to properly introduce the topic, Section 1 has given background information on the topic of XAI, formal argumentation and contrastiveness which led to the research questions of this thesis. Section 2 will provide an extensive literature research on the topic of contrastiveness within the social sciences and the humanities to conclude suitable definitions of contrastiveness. In Section 3, the multiple definitions of contrastiveness will be modelled in abstract argumentation. Then in Section 4, the same definitions will be modelled, but now in an abstract argumentation setting that considers preference relations. Section 5 steps away from abstract argumentation and focuses on modelling contrastiveness in the structured argumentation setting of ASPIC$^+$. Once this main part of the research has been done, In Section 6, I will come back to the research questions, evaluate the work that has been achieved and discuss the implications and the limitations of this research. Finally, in Section 7, the main research question will be answered, final thoughts will be shared and a direction for future research will be proposed.

1.4 Contributions

The contributions of this research are threefold. First, this thesis emphasizes the importance of XAI in general; with the rapid growth of AI applications in daily life, it is important for there to be systems that are explainable and thus users can understand, trust, and validate.

Second, this thesis contributes to the research currently done on the topic of contrastiveness. Contrastiveness is an inherent aspect of question-asking and generating explanations by humans, as has been shortly stated in Section 1.1.3 and is once again emphasised in Section 2.1. This thesis highlights the importance of contrastiveness, summarises what researchers find entails a contrastive explanation, and shows that much more research can and should be done on the topic to further improve on explainable artificial intelligence.

Thirdly and finally, this thesis also formally defines how different forms of contrastiveness (based on the relevant literature) can be modelled in argumentation - both abstract and structured. By evaluating these findings, this research's findings form a basis of how to implement contrastiveness in various XAI applications. These results can also be used as a stepping stone for further research into the topic of contrastiveness within (argumentative) XAI.

2 Contrastiveness in Literature

Contrastiveness, the question of why a certain argument (the fact) can be accepted, whilst another argument (the foil) cannot be accepted, plays a big role in explanations; Miller stated that close to *all* explanations are contrastive [Mil19]. Contrastive explanations are especially useful as they pinpoint what part of a not-understood phenomenon to explain, since foils narrow down which parts to explain. As a result, people tend to find contrastive explanations more intuitive than non-contrastive explanations, which can be far too broad [Lip90].

Considering the wide usage of contrastive explanations, it is important to know what exactly a contrastive explanation is. As was mentioned in the Section 1, however, there is no consensus on this. Much of the research done provides definitions that differ from one another. What most researchers do agree on, is that a contrastive explanation consists of explaining the contrast between the fact and foil; what are the main differences between that which did happen and that which did not ([Hes88], [Lip90], [Hil90], [MK93], [Hit99], [BW02], [Mil21], [Mil19], [TE11])? Nevertheless, there are some slight differences in how to acquire the information to explain that contrast. Below, I will go over some important notions found on contrastiveness from literature in the field of the social sciences and humanities, to see what different researchers write about contrastive explanations.

2.1 Literature Research

Hesslow was one of the earlier researchers to have written down a clear definition for contrastive explanations [Hes88]. In his paper he discussed what it is that determines the selection of the most important cause of an unexpected event. He found that this selection process cannot be arbitrary, as people tend to select explanations in similar ways to one another. If there are two situations with different outcomes, and someone wants to know why one thing happened in one situation and not in the other, then the explanandum (that to be explained) should be construed as a difference between objects with regards to a certain property, and the selection of causes should be determined by explanatory relevance. Hesslow further elaborated on this by describing the explanandum as a relation consisting of three parts: object a (the fact), the object of comparison b (the foil, which can also be a reference class R if there are multiple foils), and the explanandum property E (which are the properties that a has but the objects b in R do not have). A good explanation can be given by construing the explanandum as a difference, and thus narrowing down the many possible causes that initially existed to just those wherein the contrast lies. According to Hesslow, the highest explanatory power lies within highlighting the greatest number of differences in attributes between the target and the reference objects.

Lipton coined the terms of *the fact* and *the foil* and stated that both the fact and the foil should have a largely similar history of events, but the differences between them should stand out [Lip90]. This he called the *Difference Condition: "To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q ."* [Lip90, p256]. The contrast between the fact and the foil is explained by identifying the cause of the fact and providing the absence of the corresponding cause of the foil. This work is quite similar to the previous work of Hesslow ([Hes88]), as Lipton once again stated that contrastive explanations are easier to derive than 'complete' explanations, as a non-contrastive explanation would have to name all causes, while a contrastive explanation can narrow it down to just that wherein lies the contrast.

The findings in [Lip90] and [Hes88] are comparable, as they are rather general definitions of contrastiveness. From here on, however, the literature starts slightly deviating in defining contrastive explanations and the specifics of how such an explanation should be defined will be discussed in more detail. Hilton created his own multiple different notions of contrastive explanations [Hil90]. He stated that every *'why'* question has a *'rather than'* built in, and explanations must therefore be relevant to some question and explain why the event occurred in the target case but not in the counterfactual contrast case. In other words, he stated that all *'why'* questions are contrastive questions. The features that are shared between the target and contrast case are presupposed, and the explanatory relevance lies within the differences. Hilton related explanations more to actual conversations, and he found that all four of Grice's maxims [Gri75], which describe how people achieve effective communication in social situations, are applicable to ordinary explanation, as explanations are a form of conversation. He continued by stating that these conversational processes play a big role in explanation, and that they constrain the explanations. Building further on this, Hilton stated that there exist different types of contrast cases for contrastive explanations; different types of explanations are relevant depending on the type of question asked. Hilton defined five types of questions: the nonoccurence of an effect (*"why X rather than not X?"*), the normal case (*"why X rather than the default value for X?"*), the noncommon effect (*"why X rather than Y?"*), the prescribed case (*"why X rather than what ought to be the case?"*) and the ideal case (*"why X rather than the ideal value for X?"*). Hilton concluded that contrastive explanations are selected by these types of questions, by distinguishing between causes and conditions and that the conversational perspective on explanation constrains the explanations.

McGill and Klein also found that there is a selection process going on when people give explanations [MK93]. There exists a distinguishing factor, something that differs, that can explain an unexpected event, and explanation-focused questions are expected to generate contrastive reasoning. McGill and Klein are one of the few who start by making a clear distinction between counterfactual reasoning and contrastive reasoning. They illustrate this distinction with the example of an effect Y and a causal candidate X by asking *"Would Y have occurred if X had not?"* They concluded that in counterfactual reasoning, one focuses on instances where the candidate X is absent, while in contrastive reasoning, one considers instances in which the effect Y is absent. Different target situations were analysed, and the example they handled in their research was the question of why an employee failed. According to McGill and Klein, counterfactual reasoning is concerned with questions such as *"Would the employee have failed had she not been a woman?"*. For contrastive reasoning the question would be *"What made the difference between the employee who failed and the employees who did not fail?"* For a good contrastive explanation, there should thus be a focus on an instance in which the effect is absent, and there lies an emphasis with the *sufficiency* of a factor to give a good contrastive explanation, not specifically the *necessity* of a factor (which would be counterfactual reasoning).

Hitchcock stated again that *all* explanations convey contrastive reasoning [Hit99]. This differs from earlier research in that he says that a contrastive explanation does not require a specific type of question to be asked, but all explanations are contrastive in nature. Hitchcock then adds a new factor to take into consideration for contrastive explanations: the presupposition. Hitchcock focuses on the role that the presupposition plays while giving contrastive explanations. He defines presuppositions as a background of propositions of which all parties are aware; a sense of common knowledge. While some of his findings are mostly in line with the earlier-mentioned research ([Hes88], [Lip90] and [Hil90]) in saying that contrastive explanations explicitly contrast the explanandum (the fact) with an alternative outcome, Hitchcock also states that for something to be considered a good explanation, it needs to be directed to the part of the history we are most interested in learning about. As mentioned in Section 1.1.3, histories correspond to the events before the fact, and the events before the foil. This is where the selection process happens; presuppositions can render potential explanatory factors irrelevant, which narrows down the options for a better explanation. According to Hitchcock, explanatory relevance given a presupposition can be modelled by probabilistic relevance conditional upon the presupposition. This is in line with the findings in [Ove11] about explanations in general; Overton stated that an explanation is not just answer to a why-question, but it is the pair of answer and presupposition. Explanans a (the explanation) is the answer to why-question q with presupposition b if and only if a explains b. Thus, Hitchcock concluded that presuppositions play an important role in providing good contrastive explanations.

Van Bouwel and Weber do not agree with Hitchcock that all explanations are contrastive; there also exist 'plain fact' explanations, which show how an observed fact was caused [BW02]. Contrastive explanations, on the other hand, are defined by van Bouwel and Weber as explanations that provide information about the events that differentiate the actual causal history from its unactualised alternative. To properly explain a contrast one needs to only consider the factors that make the difference, and disregard other causal factors. Van Bouwel and Weber define multiple different kinds of contrasts: the O-contrast, where the contrast occurs between objects themselves (*"Why does object a have property P while object b has property Q?"*), the P-contrast, where the contrast occurs on properties within an object (*"why does object a have property P rather than property P'?"*), and the T-contrast, where the contrast occurs within an object over time (*"why does object a have property P at time* t_1 *but P' at time* t_2*?"*). This they called the erotetic model of explanation, and questions of these forms are meant to tell us why things have been different from what we expected them to be.

Miller stated that there exist at least two different types of contrastive why-questions [Mil21]. First, there are *Alternative Explananda*: why some fact happened rather than the foil, which are based on findings in [Lew86]. Alternative questions have the form of *"Why P rather than Q?"* Miller elaborated that in this context, fact and foil are always incompatible. This means that there is no possible manner for both the fact and the foil to be true at the same time. Secondly, Miller defines *Congruent Explananda*: why some fact happened in one situation while another fact happened in another situation, which are based on findings in [Lip90]. Congruent questions have the form of *"Why P but Q?"* In the former, the foil is hypothetical, it is something that could have happened but did not. In the latter, the foil is actual and two different events that have both occurred are compared to one another. In short, the alternative explananda ask why something gave an output rather than some other possible output, and congruent explananda ask why something gave a particular output in one situation, but a different output in another situation.

Now that various definitions for contrastiveness found in literature have been discussed, it is important to go over the properties contrastive questions and explanations should possess. This was mentioned briefly in Miller's definitions of contrastiveness with compatibility [Mil21], but it is not the only property to be defined.

First, compatibility will be discussed. In short, with regards to contrastiveness, compatibility signifies whether there exists a possibility that fact and foil are true at the same time. At first, one would assume that can not be the case, as fact and foil are often in contrast with one another (as is the case with questions in the form of *"Why P rather than not P?"*). Here it is true that fact and foil cannot both be true. However, Lipton stated that compatible contrasts do exist [Lip90]. Researchers do not agree on whether fact and foil are compatible or not - but that is due to there being different kinds of contrasts where compatibility can differ. For example, Miller stated that in alternative explananda, fact and foil are always incompatible, but he found that in congruent explananda this is not always the case [Mil21]. That is why it is important to take the property of compatiblity into consideration when defining contrastiveness.

Relevance is also an important property, if a little obvious. Fact and foil should be relevant to one another to make for a sensible question [TE11]. Relevance is defined somewhat differently across research. Lipton stated that for fact and foil to be relevant to one another, they should have a shared causal history [Lip90]. Barnes stated that this was not enough, and that both fact and foil should also be culminating events of the same causal process [Bar94]. While there is no consensus on what entails relevance specifically, in this thesis it is enough to know that fact and foil should be related to one another in one way or another, otherwise it would lead to a nonsensical question. An example of such a nonsensical question could be *"Why did Tom eat an apple rather than becoming a lawyer?"*. The contrast in this question cannot be explained, as there is no relation to Tom eating an apple and him deciding whether or not to become a lawyer. The fact and foil do not share much in their history, and thus are irrelevant to one another. Relevance is a property that all contrastive questions should adhere to.

This was an overview of the various definitions and important notions of contrastiveness and corresponding properties within the field of humanities and the social sciences. With this knowledge, I will move on towards the next section, where I will compare the commonalities between these definitions.

2.2 Reflections

What the previous section made clear, is that all research states that the highest explanatory relevance can be found in the difference between the fact and the foil ([Hes88], [Lip90], [Hil90], [MK93], [Hit99], [BW02], [Mil21], [Mil19], [TE11]). To give a good contrastive explanation, it is necessary to list the differences between the histories of the fact and the foil [Lip90]. It is important to consider at what part these histories start to differ. That is the core of a good contrastive explanation: the highest explanatory power lies within highlighting the greatest number of differences between the target and the reference objects [Hes88]. Both Lipton and Hesslow keep it at that; the difference should be noted between the two situations where there is a contrast. Contrastive explanations are not as simple as that, however, as has become evident in the previous section. Various factors can be taken into account on what kind of contrast to focus on, and an important question remains: how can the explanandum be narrowed down further?

The presupposition, introduced by Hitchcock, is a background of propositions of which all parties are aware [Hit99]. Presuppositions can render potential explanatory factors irrelevant, so it is a good mechanism to narrow down the explanandum further. Nonetheless, it can be said that this would fulfill quite a similar role to that of the foil. It is possible to view the foil as a presupposition as defined by Hitchcock, since the foil makes for more information for all parties to be aware of, and thus also narrows down to the item that is not understood well. In the example of a question such as *"Why did Mary eat the cake?"*, a foil could narrow down all the potential reasons: *"Why did Mary eat the cake instead of the potato chips?"*. Before, a possible answer could have been, 'because Mary was hungry', but the added foil 'instead of the potato chips' renders that answer irrelevant, as it does not explain why she choose one type of food above the other. The foil makes it possible that the explanation could be related to that Mary prefers sweet foods over another, or something similar. The foil creates a proposition that can render potential explanatory factors irrelevant, making the explanation more relevant, just as the presuppositions can. This is also in line with what is written in [Ove11]: a is a good explanation to a why-question, if a can explain b, with b being the foil.

In Van Bouwel and Weber's research, they wrote about different types of contrast; the P-, O- and T-contrast [BW02]. It is interesting to see how other research compares to this. Miller stated in his research that his two types of contrastive explanations, alternative and congruent, can be compared to Van Bouwel and Weber's contrasts [Mil21]. Alternative explananda, which are about why some fact happened rather than the foil, are similar to the P-contrast: the contrast occurs on properties within an object. Congruent explananda, which are about why some fact happened in one situation while another fact happened in another situation, are similar to the O-contrast, where the contrast occurs between objects themselves and to the T-contrast, where the contrast occurs within an object over time. The different situations in the congruent explananda can be either in time (T-contrast), or in the objects themselves (O-contrast).

It is also interesting to note that the different cases mentioned in [Hil90] can also be compared to the different contrasts of Van Bouwel and Weber. The ideal case and normal case, which talk about why an object would have a certain value, are comparable to the P-contrast of Van Bouwel and Weber, and thus as well very similar to Miller's alternative explananda. The noncommon and the prescribed cases can be compared to the O-contrast, as they are about the objects and not their properties. The nonoccurence of an effect, however, does not seem to have a comparable alternative in either van Bouwel and Weber, but it would be an alternative explananda in Miller's terms.

McGill and Klein state that contrastive reasoning focuses on instances where the effect is absent. A contrastive question would be *"Why did one employee fail and the other employee did not fail?"*. The effect here being 'failing', so this question displays a contrast between an employee who failed and one who did not. This contrastive reasoning can be compared to the other literature, as it is very similar to, if not exactly the same, as Hilton's nonoccurence of an effect. Thus, this type of contrast should be taken into account for contrastiveness as well, as it can be found in various work.

2.3 Defining Contrastiveness

Now that the several definitions of contrastivess in literature from the social sciences and the humanities have been compared against one another, it is time to move forward and combine these findings to create a few new definitions, based on the commonalities found in the literature. Since various notions of contrastiveness exist, and multiple researchers stated that there exists not just one type of contrastive explanation ([Hil90], [BW02], [Mil21]), to better represent the literature on contrastiveness the choice has been made to create multiple definitions to move forward with.

From the previous two subsections, it can be seen how definitions of contrastiveness can differ, but also in which ways they are similar. In summary, it can be gathered that Hilton's, Miller's, Van Bouwel and Weber's and McGill and Klein's definitions can be taken as the core of the definitions, as they can be compared well against one another and are defined quite similarly. Lipton's and Hesslow's accounts of contrastiveness are very useful, but both of them define more of a general contrastiveness and do not dive too deeply into specific types of contrasts. Miller stated himself that his two contrastive explanations are similar to the three contrasts given by Van Bouwel and Weber. Adding to that, Hilton and McGill and Klein also discuss the nonoccurence of an effect, which should be added along to that. From these findings four different types of contrasts can be derived:

Nonoccurence-contrast: the contrast occurs within the occurence and the nonoccurence of an effect (*"Why did P happen instead of not-P?"*).

Property-contrast: the contrast occurs within the properties of an object (*"Why does object a have property P rather than property Q?"*).

Object-contrast: the contrast occurs between the properties of an object (*Why does object a have property P while object b has property Q?"*).

Over Time-contrast: contrast occurs within an object over time (*"Why does object a have property P at time t_1 but Q at time t_2?"*).

Now that these four contrasts have been established, it can be studied how their properties differ from one another.

The first property to consider is relevance. This is a rather obvious property, as all types of contrast should have a relevant fact and foil. As has been elaborated upon in Section 2.1, fact and foil should have a similar causal history to be considered relevant to one another ([Bar94], [TE11]), and thus, for a contrastive question to be sensible, fact and foil should always be relevant to one another. If not, problematic questions such as the before-mentioned *"Why did Tom eat an apple rather than becoming a lawyer?"* could occur. The fact and foil do not have a similar causal history at all: eating an apple could have a cause such as hunger, while not becoming a lawyer could have a cause such as not being able to pay for tuition. Thus, to adhere to the relevance property, fact and foil should have a similar causal history for a contrastive question to be sensible. This is the case for all four types of contrast.

Secondly, we will discuss compatibility. As mentioned in Ssection 2.1, compatibility regards whether fact and foil can be true at the same time. Some research states that by definition of a contrast, fact and foil should not be able to be true at the same time [Gar82]. Some others have stated that compatible contrasts do exist [Lip90] [Bar94]. It is important to note that this differs per type of contrast, which can be backed up by literature.

Starting with the nonoccurence-contrast, it will be quite obvious to see why in this case fact and foil cannot be compatible by definition. Nonoccurence-contrasts consider the occurrence and nonoccurrence of the same event, and it is simply impossible for something to both happen and not happen. Thus, for nonoccurence-contrasts, fact and foil are never compatible.

The next contrast is the property-contrast. The situation is comparable to that of the nonoccurence-contrast, as fact and foil also cannot be compatible in this case. Property-contrasts consider contrasts within an object, and the foil is an alternative possibility to the fact [Mil21]. Taking the example of the question: *"Why is this leaf blue rather than yellow?"*, it is impossible for the leaf to be both blue and yellow at the same time. It is worth mentioning that it is possible for some yellow leaf to exist, or some leaf that is both blue and yellow, but that is not what is questioned here; the explainee is asking why *this particular leaf* is blue. Fact and foil refer to the same object and thus, for property-contrasts, fact and foil are incompatible.

Next is the object-contrast, where the situation changes and it is possible for fact and foil to be compatible. An example illustrated by [TE11] is that of the question *"Why did Willie Sutton rob banks rather than gas stations?"*. In this case, the contrast is compatible because it was possible for Willie to rob both banks and gas stations throughout his career in crime [TE11]. In this case, the foil is something that can actually happen, or happened in another situation, and is called a surrogate [Mil21].

And finally, the over time-contrast. In this case, fact and foil cannot be compatible. As was stated in an example in [DB08]: *"Why did Earth develop life after one billion years rather than after one million years?"*, it it impossible for both fact and foil to be true in this case. If life developed at an earlier state, it is impossible that it *also* developed later. And if it developed later, it cannot *also* have already happened earlier. Thus, for over time-contrasts, fact and foil cannot be compatible.

The overview of these properties (relevance and compatibility) can be found in Table 2.3.

An observant reader might have noticed that three of the contrasts are comparable to the contrasts found in the earlier research [BB22b]. I am not the first to try and create definitions for contrastive explanations based on research from the social sciences and the humanities. In their research, Borg and

Property	**Nonoccurence-Contrast**	**Property-Contrast**	**Object-Contrast**	**Over Time-Contrast**
Relevance	yes	yes	yes	yes
Compatibility	no	no	yes	no

Table 2: Overview of properties per type of contrast

Bex also gave multiple definitions for contrastiveness within an abstract argumentation setting, based on literature from the social sciences [BB22b]. They defined three distinct definitions: the negation question (*"why P rather than not-P?"*), the property question (*"why P rather than the default value for P?" or "why does object a have property P, rather than property Q?"*), and the object question (*"why does object a have property P, while object b has property Q?"*). As cited in their research, their definitions are mostly built on the before-mentioned research [Hil90] and [BW02]. It is possible to compare my nonoccurence with their negation-contrast, my property with their within and despite-contrast and my object with their between and despite-contrast. By doing my own research, my concluding findings on the topic of types of contrasts are comparable to theirs. This implies that their findings are very strong and grounded, and I agree with their definitions of constrastiveness. The main difference is that my findings added an over time-contrast. The details of the similarities and differences of the contrasts will be discussed in Section 3.4.

2.4 Requirements for Contrastiveness

Aside from the properties, there are also some other requirements that should be met for both a sensible contrastive question and a sensible contrastive explanation. Sensible meaning that the question and explanation should make sense, there is a certain logic to them. To make sure my research complies with such requirements, I will go over those that I found to be most important and relevant. First, requirements for contrastive questions will be discussed. These requirements are based on findings by [TE11], [Bar94] and [DB08].

The first requirement is that of the truth value. For a good contrastive question to be sensible, it is necessary that the fact and the negation of the foil are both true statements [TE11]. For example, taking the earlier mentioned example question for a property-contrast, *"Why is this leaf blue rather than yellow?"* it must be the case that the leaf is blue, and not yellow for the question to be sensible. While it is possible for yellow leaves or multicoloured leaves to exist, that is not what is questioned here; if the leaf in question is actually yellow, there is no point in asking this question, and no adequate explanation can be given. Thus, for a question to be sensible, it is necessary that P is true, and the negation of Q is also true.

Secondly, the type of contrast needs be determined. To know what kind of contrast will be discussed (nonoccurence, property, object or over time), it is important that the contrast is made explicit. As has been mentioned in earlier sections, when people ask for an explanation, sometimes the foil is left implicit. When asked *"Why did Mary eat the cake?"*, there are a multitude of possible explanations, while generally only one specific one is wanted. This makes for an incomplete contrastive question, since it has not yet been determined what type of contrast is meant. There needs to be a way to find the contrast, even when one is not explicitly given. This can be achieved in the setting of formal argumentation. This is due to formal argumentation being able to derive a foil when none is provided, since it comes with an explicit notion of conflict [BB21b]. Thus this way, the type of contrast can be determined.

Finally, the difference to be explained should point towards only a single contrast [DB08]. This is illustrated in the earlier example of Tom eating an apple rather than becoming a lawyer - two different aspects are referred to: the possibility of what to eat and the possibility of what career to follow. It is uncertain which contrast is meant as the question refers to two unrelated aspects. As a result, the fact and foil are not relevant to one another, and thus the question becomes nonsensical.

After a sensible contrastive question has been asked, a contrastive explanation is what follows. Aside from the type of contrast that there is to consider for a contrastive explanation, there are also different values that can be used to measure how good, or 'powerful' a contrastive explanation is. The measures I will consider for my research are based on the findings by Ylikoski and Kuorikoski [YK10], and they can be used to compare, evaluate, and distinguish the goodness of my definitions of contrastiveness. These measures can attribute explanatory power to an explanation, which can then be used to compare explanations against one another, and the more 'powerful' explanation is the preferred one. This can help us get insight into which explanation out of multiple would be preferred, and give us insight into the reasons as to why this is. Thus, the important measures to determine explanatory power given by Ylikoski and Kuorikoski will be discussed [YK10].

The first measure is non-sensitivity. If an explanation is sensitive, it means that the explanatory power of the explanation is dependant on changes within the background conditions. This is undesirable, as one would prefer their explanations to be stable and still work when, for example, just a single variable is changed in the question. In short, the more sensitive an explanation is, the less powerful it is and as a result, non-sensitivity is preferred. It is important to note, however, that overall sensitivity is irrelevant since not all background conditions are equally important.

The second measure is precision, regarding how precisely the explanation characterises the explanandum. Generally, the more detailed an explanation is, the better it is. There does exist a trade-off between non-sensitivity and precision, however. The sensitivity of an explanation usually increases when precision of the explanandum is increased.

The third measure is factual accuracy. At first, it would be assumed that this measure rather speaks for itself; explanations should be factual, as a false explanation cannot possibly be a good explanation. It is important to consider that all of this is relative however. In some cases, it is not possible for an explanation to be entirely factual, and in such cases the explanation with the highest factual accuracy out of all of them is the one to be preferred and the one with higher explanatory power.

The fourth measure is the degree of integration. An explanation is considered to be more credible if it is consistent with a well-supported theory. An integrated body of knowledge is more than the sum of its parts, and people would prefer that over an explanation that is based on nothing conjured out of thin air. Thus, a higher degree of integration makes for a more powerful explanation.

The fifth and final measure is cognitive salience. Cognitive salience comprises of the ease with which the reasoning of an explanation can be followed. Explanations that consist of terminology or argument structures that one is familiar with are more cognitively salient than explanations that use widely unfamiliar terms. Cognitive salience does differ from simplicity, as cognitive salience is more connected to the cognitive capabilities of a person, which can differ greatly, while simplicity is far more general.

These five measures can be used to compare multiple possible explanations against one another, and can provide insight into why one could be considered the better and more powerful explanation over another. Especially with contrastive explanations, where I have defined different possible contrasts to answer a question with, it is useful to be to able use these measures to compare why one contrast could possible give an explanation that is preferred over another.

2.5 Overview of Results Thus Far

This subsection provides a succinct overview of the gathered results thus far in Section 2. First, from a thorough overview of literature it was concluded that there exist four different types of contrasts: the nonoccurence-contrast, the property-contrast, the object-contrast and the over time-contrast. These are the four definitions of contrastiveness to move forward with. Then, three requirements were decided: that of truth value, the type of contrast and that there must only be one contrast in a question. All contrastive questions should abide by these requirements. Finally, five measures of how good an explanation is were decided: non-sensitivity, precision, factual accuracy, degree of integration and cognitive salience. These measures will be used in Section 6.2 to evaluate my contrastive explanations.

3 Contrastiveness in Abstract Argumentation

Now that the various concepts of contrastiveness have been discussed, the next step is to formalise them in an abstract argumentation setting. Before this can be done, it is important to discuss some preliminaries. Thus, in this section, first the preliminaries of abstract explanations will be elaborated upon, after which contrastiveness will be formalised in an abstract argumentation setting.

3.1 Abstract Argumentation Preliminaries

Argumentation theory is the study of how conclusions can be supported or undermined by premises through logical reasoning. It is a highly interdisciplinary field with links to psychology, linguistics, philosophy, legal theory, and formal logic [CDG$^+$15]. Argumentation is a useful tool to decide for or against certain claims, and it allows for explanations as to why a certain argument or claim can be accepted. Argumentation is thus suitable for explanation selection [Mil19].

In this section, the focus will be on abstract argumentation. Popularised by Dung ([Dun95]), abstract argumentation is based on the fundamental mechanism humans use in argumentation. A particular feature is the simple structure of the frameworks used: in abstract argumentation, all that is represented are the sets of arguments and the attack relations between them; it does not specify the internal structure of the arguments. The arguments represent all that can be argued for. Inconsistencies within the knowledge base are represented by conflicts among arguments, modelled via directed attacks between arguments. A benefit of abstract argumentation is that systems for argumentation-based inference can handle incomplete, inconsistent, and uncertain information [Pra18]. This reasoning is known as non-monotonic logic. Non-monotonic logic is a form of reasoning that allows for the handling of defeasible reasoning, where conclusions can be revised or defeated based on conflicting arguments, which allows for a more flexible approach to reasoning in complex and uncertain situations. In terms of argumentation, it means that it allows for the revision of conclusions in the presence of conflicting evidence, which allows one to decide on certain conclusions as long as nothing proves otherwise.

Abstract argumentation studies Argumentation Frameworks (AFs), which is a pair of abstract arguments and binary attack relations between those arguments [Dun95].

Definition 3.1. An *Argumentation Framework (AF)* is a pair $(Args, att)$, where $Args$ is a finite set of arguments, with binary attack relations on the arguments $att \subseteq Args \times Args$.

We say that argument A attacks argument B if and only if $(A, B) \in att$. We also say that argument A defends argument B if A attacks the attacker of B. An argument A indirectly attacks another argument B if A attacks an argument that defends B. We say that an argument A indirectly defends another argument B if A attacks an argument that indirectly attacks B. Similarly, we say that a set S of arguments (indirectly) attacks B if B is (indirectly) attacked by an argument in S. A set S of arguments (indirectly) defends B if B is (indirectly) defended by an argument in S [Dun95]. It is of importance to note that an argument always defends itself [FT15]. Let us illustrate this with a small example, as all argumentation frameworks can be represented as a directed graph.

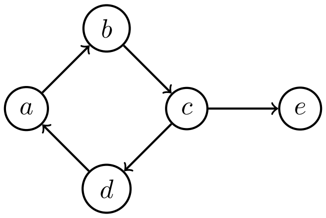

Figure 1: an example abstract argumentation framework

Example 1. *Figure 1 shows a simple example of an AF, with* $Args = \{a, b, c, d, e\}$ *and* $att = \{(a, b), (b, c), (c, d), (d, a), (c, e)\}$.

Dung's abstract argumentation frameworks have been chosen to use in this research, as they are the most simple and straightforward. Other abstract argumentation frameworks do exist (Bipolar Argumentation [CLS05], Support Argumentation [CLS05], Quantitative Bipolar Argumentation [BRT19], among others), but all of these are built upon Dung's ideas. Dung's abstract argumentation frameworks form the basis of these other approaches, and thus it is appropriate to start with that - and afterwards expand upon the findings in this section in other frameworks, as has been done with preference-based argumentation in Section 4 and ASPIC$^+$ in Section 5.

Aside from introducing the abstract argumentation framework, Dung also addressed how different sets of arguments can be accepted together. An argument A is acceptable with respect to a set of arguments S, if the argument can be defended against all its attackers [Dun95].

Definition 3.2. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and $S \subseteq Args$ be a set of arguments. Then: An argument $A \in Args$ is acceptable with respect to S if and only if for each argument $B \in Args$: if B attacks A then B is attacked by S.

In abstract argumentation, an extension refers to a set of arguments that is considered collectively acceptable within a given argumentation framework. Extensions have no internal conflicts, which means there are no attack relations between the arguments. Different extensions exist with different criteria for the arguments, such as meeting specific rules of acceptance or defeating opposing arguments. Applying those different criteria can lead to different extensions, and a noteworthy type are the admissible extensions, which are non-conflicting and counter-attack any attack from outside, providing a way to argumentatively defend each argument within the extension [Dun95].

Definition 3.3. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and $E \subseteq Args$ be an extension. Then: E is conflict-free if there are no $A, B \in E$ such that $(A, B) \in Att$; E is admissible if it is conflict-free and it defends all of its elements.

Example 2. *In the case of the AF displayed in Figure 1, the admissible extensions are* $\{\}, \{a, c\}$, $\{b, d\}$ *and* $\{b, d, e\}$.

Acceptability with regards to a set S has been defined in Definition 3.2, but the acceptability of an argument has yet to be defined:

Definition 3.4. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and let $A \in Args$ be an argument. $Adm(\mathcal{AF})$ denotes the set of all the admissible extensions of $\mathcal{AF}$. Then: A is accepted if there is some $E \in Adm(\mathcal{AF})$ such that $A \in E$; and A is not accepted if there is no $E \in Adm(\mathcal{AF})$ such that $A \in E$.

Definition 3.4 states that an argument A is accepted, if there exists an admissible extension of which it is a part. If there is no admissible extension of which A is a part, then the argument A is not accepted.

3.2 Explanations in Abstract Argumentation

As has been mentioned in Section 1.1.2, argumentation is suitable for explanation in XAI. Arguing for an argument can be deemed as an explanation, as argumentation is capable of transparently explaining the procedure and the results of reasoning [FT15]. Following this mindset, an explanation can be modelled as a justification of the explanandum. In this section, such a justification will be defined as sets of arguments, to use as explanations. Specifically, sets of arguments in an abstract argumentation framework.

Admissible extensions, which are defined in Definition 3.3, can be used as a justification and thus as an explanation for an argument. Admissible extensions will be used as a basis for explanations in this research. If an argument belongs to an admissible extension, that extension can be used as an explanation for that argument. If an argument does not belong to any admissible extension, it means that other arguments will always defeat it. Thus, there is no possible manner to justify the argument and therefore the acceptance of that argument cannot be explained either. Based on this idea explanations can be formally introduced:

Definition 3.5. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF. First, let $P \in Args$ be accepted, then:

$$\mathsf{In}(P) = \{S \mid S \text{ is an admissible extension and } P \in S\}.$$

Now let $Q \in Args$ be non-accepted, then:

$$\mathsf{Out}(Q) = \{S \mid S \text{ is an admissible extension, } Q \notin S,\ P \in S \text{ and } P \times Q \in Att\}.$$

$\mathsf{In}(P)$ denotes the set of all the admissible extensions of which P is a part as possible explanations for P, while $\mathsf{Out}(P)$ denotes the set of all the admissible extensions of which P is not a part, which attack P, as an explanation for *not* P, the non-acceptance of P. While empty sets can be admissible extensions, this makes it so that empty sets are not included for explanation; only when there is no viable explanations at all.

Example 3. *In the case of the AF in Figure 1, $\mathsf{In}(a)$ would give the admissible extension $\{a, c\}$, while $\mathsf{Out}(a)$ would give the two admissible extensions $\{b, d\}$ and $\{b, d, e\}$.*

This notation implies that the non-acceptance of an argument can be denoted by the negation of P ($\neg P$). In other words, $\neg P$ can be explained by using $\mathsf{Out}(P)$, as $\mathsf{Out}(P)$ gives the justification for why argument P is not accepted, which is $\neg P$. So, from this point on in this research, when a question of the form *"Why $\neg P$?"* is asked, possible explanations can be given by $\mathsf{Out}(P)$.

3.2.1 Necessity and Sufficiency in Abstract Argumentation

By using admissible extensions in abstract argumentation as possible explanations, one can be left with large sets of arguments as explanations in the case of larger argumentation frameworks. For example, consider an argumentation framework with 100 arguments and no attack relations: all 100 arguments together form an admissible extension (as it is conflict-free and defends itself), but it does not provide a clear explanation of why one of those arguments is accepted. A selection mechanism is required, to narrow down the possible explanations. Two selection mechanisms that exist are necessity and sufficiency ([Lip90] and [Lom10]).

Necessity and sufficiency can be illustrated at the hand of an example in natural language. Having four sides is a necessary, but not a sufficient condition for something to be square. On the other hand, something being square is a sufficient, but not a necessary condition for it having four sides. Rectangles have four sides, but are not square. In short, A is necessary for B, if in order for B to happen, A has to happen as well. A is sufficient for B if no other causes are required for B to happen. Translating this idea to explanation within argumentation, a necessary explanation contains the arguments that one has to accept in order to accept the considered argument, while a sufficient explanation contains the arguments that, when accepted, guarantee the acceptance of the considered argument [BB21c].

Example 4. *Imagine an AF with $Args = \{a,\ b,\ c,\ d,\ e,\ f\}$, and $Att = \{(b, a),\ (c, b),\ (c, d),\ (d, c),\ (e, b),\ (e, f),\ (f, e)\}$ (see Figure 2, a variation of the framework in [BB21c]). For a to be accepted, b has to be defeated. This is the case if c, e or c and e are accepted. Both c and e are not necessary arguments for a, as neither argument has to be accepted to accept a: one or the other will do. However, both arguments are sufficient.*

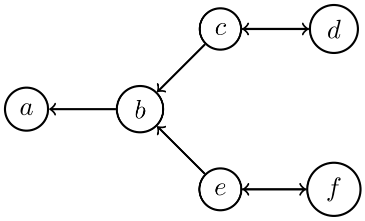

Figure 2: an example abstract argumentation framework with independent attacks (simplified variation of the framework in [BB21c]

It has been argued that necessity and sufficiency are strong criteria for preferred explanatory causes [Mil19]. Lipton stated that necessary causes are preferred to sufficient causes [Lip90]. Necessity, however, cannot keep up with multiple separately sufficient causes. As a result, Woodward argued that sufficiency is a stronger criteria, as it can bring about an effect without any other causes [Woo06]. The same can be said in an abstract argumentation setting about necessity and sufficiency. Necessity cannot always provide explanations, as it cannot account for two separate, independent attacks on one argument. In such a case, either argument would be sufficient for the non-acceptance of the argument, but neither argument is strictly necessary (see Example 4). Thus, to further select explanations, sufficiency will be used: what (sets of) arguments are sufficient for the (non-)acceptance of an argument?

To find the sufficient arguments of an explanation, a new definition for explanations in abstract argumentation will be necessary. This definition will be based on the definition for sufficiency in [BB21c]:

Definition 3.6. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and $P \in Args$. Then:

$S \subseteq Args$ is sufficient for the acceptance of P if S is conflict-free and all $A \in S$ (indirectly) defend P against its attackers.

$S \subseteq Args$ is sufficient for the non-acceptance of P if S is conflict-free, all $A \in S$ (indirectly) attack P and all defenders of P are (indirectly) attacked by an $A \in S$.

Definition 3.6 states that a set is only sufficient for P if all of its arguments either (indirectly) defend P (for acceptance), or (indirectly) attack P and all its defenders (for non-acceptance). By including that *all* arguments either (indirectly) attack or defend the argument, only truly relevant arguments will be considered as sufficient for the (non-)acceptance of an argument. Based on this definition of sufficiency, a new definition will be created to find the sufficient explanation for the (non-)acceptance of an argument:

Definition 3.7. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF. First, let $P \in Args$ be accepted in an admissible extension, then:

$$\mathsf{SuffIn}(P) = \{S \subseteq Args \mid S \text{ is an admissible extension, } P \in S,\ S \text{ is sufficient for the acceptance of } P\}.$$

Now let $Q \in Args$ be non-accepted in an admissible extension, then:

$$\mathsf{SuffOut}(Q) = \{S \subseteq Args \mid S \text{ is an admissible extension, } S \text{ is sufficient for the non-acceptance of } P\}.$$

In natural language, Definition 3.7 states that $\mathsf{SuffIn}(P)$ denotes the sufficient admissible extensions of which P is a part, and of which all the arguments (indirectly) defend P against all its attackers as possible explanations. $\mathsf{SuffOut}(P)$ denotes the sufficient admissible extensions of which all the arguments (indirectly) attack P as possible explanations. The addition of sufficiency differentiates this definition from Definition 3.5. Another difference with Definition 3.5 is that in the case of a sufficient explanation, all the arguments in the admissible extension must (indirectly) *defend* or *attack* the argument. In this manner, a notion of relevance is modelled, and irrelevant arguments will not be included in the explanation.

Example 5. *In the case of the AF in Figure 1, $\mathsf{SuffIn}(b)$ would give the set $\{b, d\}$. This differs from $\mathsf{In}(b)$, which would give the two sets $\{b, d\}$ and $\{b, d, e\}$. While the latter is an admissible extension, argument e does not (indirectly) defend b. This means that e is not a sufficient argument for the acceptance of b, and should be excluded from the explanation for the acceptance of b.*

Example 6. *In the case of the AF in Figure 2, $\mathsf{SuffOut}(a)$ would give the sets $\{\{d, f\}, \{b, d, f\}\}$. This differs from $\mathsf{Out}(a)$, which would give the set $\{b, d, f\}$. $\mathsf{SuffIn}(a)$ would give the sets $\{\{a, c, e\}, \{a, c\}, \{a, e\}\}$. In both cases, there are multiple possible explanations for why argument a could be (non-)accepted.*

As can be gathered from Example 6, $\mathsf{SuffIn}(P)$ and $\mathsf{SuffOut}(P)$ can return multiple sets of arguments. This is perfectly valid, as it is possible that there would be multiple sufficient explanations.

By taking sufficient arguments into account, the explanation will be narrowed down more by only considering arguments that are truly relevant to the explanandum. This concept of sufficient admissible extensions is very similar to initial sets, coined by Xu and Cayrol [XC18]. They stated that admissibility alone is not sufficient to model explainability as it does not take relevance into account. They solved this by considering minimal (with regards to set inclusion) admissible extensions, which they called initial sets. The difference, however, is that my sufficient admissible extensions are sufficient with regards to the (non-)acceptance of a specific argument, and are thus better suited for explanations about the (non-)acceptance of such an argument.

3.2.2 Empty Explanations

A problem that might occur is that of empty explanations: when asking for an explanation for the (non-)acceptance of an argument, an empty set could be given, which explains nothing to the explainee. Empty explanations will only be the case if, for acceptance of an argument, that argument is not part of any admissible extension, or for non-acceptance, that argument is part of every possible admissible extension.

As stated before, an argument always defends itself [FT15], and as a result, it is more likely one might end up with the singular argument one is attempting to explain as its sole justification, rather than a completely empty explanation. While this is not desirable, occasionally this might just be the case; it is possible that an argument cannot be explained further. In the case of an argument with no attack relations - unconnected to the rest of the framework - that argument itself will be its sole justification, and thus explanation. In very small abstract argumentation frameworks, with few attack relations, this might also be the case (see Example 7).

Figure 3: an example abstract argumentation framework with a two arguments and no attack relations

Example 7. *Consider the AF in Figure 3. $\mathsf{SuffIn}(a)$ would result in the set $\{a\}$, with the argument a being the sole justification for its own acceptance, as an argument always defends itself. $\mathsf{SuffOut}(a)$ would result in the empty set $\{\emptyset\}$, as there are no sufficient arguments for the non-acceptance of a.*

3.3 Contrastiveness

Now that the preliminaries of explanations in abstract argumentation have been discussed, it is time to model the concept of contrastiveness in abstract argumentation. Before we turn to the four contrasts, some properties and requirements of contrastiveness will be discussed.

3.3.1 Properties

As has been mentioned in Section 2.3, one of the properties of contrastiveness is relevance: both fact and foil should be relevant to one another. Now that contrastiveness will be defined in an abstract argumentation setting, it is important that this relevance is modelled as well. This will be done in the same manner as in [BB22b]. Borg and Bex modelled relevance with the help of attack relations: attack relations between fact and foil account for relevance. So, for fact P and foil Q to be relevant to one another, either P should attack Q, or Q should attack P. It is important to note that indirect attack relations account for relevance as well. This could be the case for situations such as an argument R defending Q, and P attacking R. In this case, P indirectly attacks Q, and as a result, P and Q are still considered relevant to one another. If there is no attack relation at all, direct or indirect, fact and foil are not relevant to one another, and thus cannot make for a sensible contrastive question. This way of modelling relevance is simple and succinct, and is worthwhile to continue working with as has been shown in [BB22b].

The second property mentioned in Section 2.3 is compatibility. Based on literature research in the field of humanities and the social sciences, it can be agreed that in some cases of contrastiveness, fact and foil *can* be compatible, but not in all ([Lip90], [Yli07]). To define compatibility in an abstract argumentation setting, we can look at the admissible extensions the fact and foil are both accepted in. Only if the fact and the foil can be accepted in the same admissible extension, are they considered compatible. If no such extension exists, fact and foil are considered incompatible. Thus, the four contrasts should be defined in such a way that only the object-contrast has admissible extensions in which both fact and foil can be accepted, and the other three contrasts (nonoccurence-contrast, property-contrast and over time-contrast) do not have any extensions in which fact and foil are compatible. This manner of modelling compatibility is also very similar to earlier work done in [BB22b]. Compatibility depends on the type of contrast in the question - and compatibility is only the case when two different situations are compared against one another in the contrastive questions.

3.3.2 Requirements

The requirements mentioned in Section 2.4 should also be maintained. Abstract argumentation is capable of adopting these requirements, and this subsection elaborates on how that is done.

The first requirement is that the fact and the negation of the foil should be true. Modelling this in abstract argumentation would entail that for the fact there should exist at least one admissible extension of which it is a part, and for the foil there should exist at least one admissible extension of which it is *not* a part. One could not ask why argument P is accepted rather than not, when there exists no admissible extension where P is accepted. In such a case, no explanation could be given.

Secondly, the type of contrast should be determined. For the rest of this thesis, the following notions will be adopted: a contrast is a nonoccurence-contrast when the foil is the negation of the fact, a contrast is property-contrast if fact and foil are different arguments and incompatible, a contrast is an object-contrast if fact and foil are compared between two different situations, and possibly are compatible, and an over time-contrast if the fact and foil are compared over a changing situation. Once the type of contrast is determined, the explanation can be found by applying the right formal definition of that contrast.

Thirdly, only one type of contrast should be used. That is why for each type of contrast, there will be one specific formalisation for abstract argumentation, and it is be determined which contrast is the case when a contrastive question is asked.

3.3.3 Modelling Contrastiveness

The modelling of the four different concepts of contrastiveness will be discussed in this subsection: the nonoccurence-contrast, the object-contrast, the property-contrast and the over time-contrast. To model the definitions of contrastiveness in an abstract argumentation setting, a slightly modified version of the earlier example framework will be used, see Figure 4. By adding one argument with two attack relations, the amount of admissible extensions has gone up.

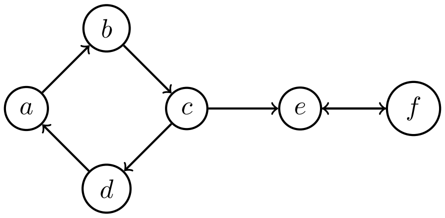

Figure 4: slightly modified example of the earlier abstract argumentation framework

Example 8. *In the case of the AF displayed in Figure 4, the admissible extensions are* $\{\}$, $\{a, c, f\}$, $\{a, c\}$, $\{b, d, e\}$, $\{b, d, f\}$, $\{b, d\}$ *and* $\{f\}$.

All the possible contrasts will be elaborated upon with examples taken from Figure 4. The first contrast to define is the nonoccurrence-contrast, where the questions that are asked are of the form *"Why P rather than not P?"*. The contrast occurs by having the foil be the negation of the fact. By definition, this makes the fact and foil incompatible; there will never exist extensions where one argument is both accepted and not accepted at the same time. By specifying that the explanation should focus on the contrast between P and $\neg P$, the explanation will differ from other contrasts.

Definition 3.8. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and let $P \in Args$, then:

$$\mathsf{NonCont}(P) = \mathsf{SuffIn}(P).$$

The definition of the nonoccurence-contrast finds the sufficient admissible extensions in which P is accepted. As mentioned in 3.2.1, A is sufficient for P if no other causes are required for P to happen. Thus, by finding the sufficient arguments for P, P will happen, which explains why P and immediately implies why $\neg P$ as well. Sufficient arguments for P will never include sufficient arguments for $\neg P$, thus just finding the sufficient arguments for P will be enough to explain the contrast.

Example 9. *Let us look at the AF Figure 4 and wonder "Why a rather than not a?". The acceptance of a has one sufficient explanation:* $\{a, c\}$*, Thus, the explanation for a is* $\{a, c\}$*.*

Moving on towards the property-contrast, we encounter questions of the form *"Why P rather than Q?"*. The situation changes slightly. Now the fact and foil are two different arguments: the fact should be accepted and the foil should not be accepted. Additionally, there should exist an (indirect) attack relation between the fact and the foil for them to be relevant to one another. Since fact and foil are not compatible in this case, there should exist no admissible extensions where both fact and foil are true at the same time.

Definition 3.9. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and let $P,\ Q \in Args$, then:

$$\mathsf{PropCont}(P, Q) = \left\{ S \setminus \bigcup \mathsf{SuffIn}(Q) \mid S \in \mathsf{SuffIn}(P) \right\}.$$

For all possible explanations for P, this definition takes the set difference with all the arguments that can sufficiently explain the acceptance of Q. This way, it returns again a set of possible explanations for *"Why P rather than Q?"*. The explanations would consist of the sufficient arguments for the acceptance of P, which are not part of the sufficient sets of arguments for the acceptance of Q.

Example 10. *Let us look again at Figure 4 and wonder "Why b rather than c?". The acceptance of b has the admissible extensions* $\{b, d\}$ *as the only sufficient explanation. The sufficient arguments for c are* $\{a, c\}$*. Taking the set exclusion between these two explanations, one is left with* $\{b, d\}$ *as an explanation for b.*

The third contrast to define is the object-contrast. The object-contrast considers questions such as *"Why P and not Q, rather than P and Q?"*. As can be seen in Table 2.3, the object-contrast is the first and only case where both fact and foil can be true at the same time. The focus of this type of questions is the difference between two situations; one where the fact is true and the foil is not, and an alternate possibility where both are true at the same time. Thus, the object-contrast can only be used when there exists an admissible extension in which both the fact and the foil are true.

Definition 3.10. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and let $P,\ Q \in Args$, then:

$$\mathsf{ObjCont}(P, Q) = \left\{ S \setminus \bigcup (\mathsf{SuffIn}(P) \cap \mathsf{SuffIn}(Q)) \mid S \in (\mathsf{SuffIn}(P) \cap \mathsf{SuffOut}(Q)) \right\}.$$

To answer a question such as *"Why is P true and not Q, rather than P and Q being true?"*, where fact and foil do not necessarily exclude one another, it is important to make it clear that two different situations are compared: one where the fact is true but the foil is not, and one where both are true. Thus, for the first situation one finds the sufficient arguments for why P is true and Q is not. The next step is to find the sufficient arguments for why P *and* Q are true. Finally, the sufficient arguments for P *and* Q should be excluded from the sufficient arguments for P and *not* Q. This is done for every possible explanation for P and not Q. This will leave the explanations for why P is true and Q is not in this specific situation, even though there exists a possibility for both to be true.

Example 11. *Continuing with the same example framework of Figure 4, we now ask the question "Why is b true and not f, rather than b and f being true?". The first step is to find the sufficient arguments for b and not f,* $(\mathsf{SuffIn}(b) \cap \mathsf{SuffOut}(f))$*, and the only admissible extension in which b is true and f is attacked, is the extension* $\{b, d, e\}$*. The next step is to find the sufficient admissible extensions in which b and f are both true, which in the case of this example is the extension* $\{b, d, f\}$*. Taking the set difference of* $\{b, d, e\} \setminus \{b, d, f\}$*, leaves us with the argument* $\{e\}$*. Thus,* $\{e\}$ *is the sufficient explanation for the question "Why is b true and not f, rather than b and f being true?", as the difference in those two situations is the acceptance of e.*

And finally, the last contrast is the over time-contrast. This contrast is concerned with questions such as *"Why P at time t_1 rather than at t_2?"*.

So far, all contrasts have been successfully modelled in a static argumentation frameworks: the framework itself is set and does not change. However, in practice, argumentation is a dynamic process in which not all arguments may be known in advance. To model the over time-contrast, one needs to model a change in time in the same situation: a change in the framework must be modelled. Different ideas have been proposed to handle such dynamics and one example is the Incomplete Argumentation Frameworks (IAFs) ([CDLS07] [BNRS18] [BJN+21] [MR20]). In an IAF, arguments and attacks are split into two parts: a certain and an uncertain part. The uncertain parts can be added to the framework at a later point, to model the obtaining of new information, or removed when finding that the element does not fit with the setting [OBB+22].

Studying contrastiveness within IAFs, however, is outside of the scope of this research. Nonetheless, it would be a worthwhile topic to dive into for future research. For this research, from this point on we will be moving forward with just the other three contrasts, which can be modelled in static argumentation frameworks.

3.4 Comparing Related Research

An attentive reader might have noticed similarities between this research and past work on similar topics. Indeed, the findings in this research so far can be compared against the findings in [BB22b], as both research is concerned with modelling contrastiveness in argumentation (specifically, abstract argumentation). In this section, the similarities and differences between related research on the topic will be discussed and evaluated against one another. To the best of my knowledge, this related research consists of the works [BB22a] and [BB22b].

3.4.1 Differences

While there are similarities in our findings, there are also some differing conclusions. This section will go over what differs between this research and existing research.

The first main difference is the usage of admissible extensions. In this research, an admissible extension can be used as a justification, and thus as an explanation, for an argument. This is based on findings in [FT15] and [UW21]. In [BB22b], their explanations are based on different extensions: grounded, complete and preferred. A complete extension is an admissible extension that contains all arguments it defends, a grounded extension is a minimal (with respect to set inclusion) complete extension and a preferred extension is a maximal complete extension. Their explanations are grounded in these extensions, and a single explanation will be given for an argument in a specified extension. In [BB22a], admissible extensions are also included. Once again, however, the specific extension is specified in which the explanation will be given. My work is broader: giving the possibility of multiple explanations for why an argument could be accepted in general, not just in a specific extension. By specifying an extension, one is only looking at one possible setting, while my definitions give explanations for all possibilities of acceptance of an argument, not limited to a specific extension, which can result in a more certain explanation, and more detailed - as it considers all possibilities.

A second difference is the use of sufficiency. This was a choice made based on the usage of admissible extensions and not specifying one extension: as a result, a wide range of possible explanations can be given for an argument. Necessity and sufficiency can be used to narrow down explanations, which is also one of the goals of contrastiveness. Necessity, however, is too ‘strict’, it filters out too many arguments in an explanation and it cannot handle explanations with independent attacks. Thus, it was decided to implement sufficiency. Borg and Bex did not make use of sufficiency in their research on contrastiveness in abstract argumentation ([BB22a], [BB22b]). This decision makes sense for them, as their explanations are given in one specific extension. Coincidentally, the manner in which this work implemented sufficiency makes it so that the definition for sufficient explanations becomes very similar to their definition of relevance. In Section 3.2.1, sufficient admissible extensions were defined as *admissible extensions of which all the arguments (indirectly) defend P against all its attackers.* The part of all arguments (indirectly) defending P, implies the relevance noted in [BB22b]. Through different paths, comparable results were found. It is also of note that in [BB21c] they did discuss necessary and sufficient explanations for abstract argumentation. The idea of using sufficient arguments in abstract argumentation is thus not new but based on earlier findings in [BB21c].

A third difference is that in [BB22b], Borg and Bex gave more possible combinations of types of contrasts, making a distinction between *‘given’* and *‘despite’* type contrasts, and *‘within’* and *‘between’* type contrasts. The ‘within’-contrasts are about questions asking for a contrast within the same extension, and the ‘between’-contrasts are about contrasts between two different extensions. For questions of the form *“Why P given Q?”*, the explanations should contain what fact and foil have in common. For the questions of the form *“Why P despite Q?”*, the explanation for the fact does not take into account the explanation for the foil. This is a distinction I implemented in another way: in this work I did not assume a different 'between' possibility per contrast, but defined the object-contrast is such a way that the admissible extensions are used in which the fact is accepted and foil is not, and the difference is taken with the admissible extension(s) in which fact and foil are both accepted. In this way, it is still possible to compare various admissible extensions, while not just being limited to the two specified - but any in which the requirements of acceptability of fact and foil hold. The 'within'-contrast I did not account for, as in all cases all possible admissible extensions are used, to provide the explanations in all possible scenarios. Never is one admissible extension specified in this work. 'Despite' and 'given' questions are also not something I accounted for in this research, rather focusing only on 'rather than' type of contrastive questions.

3.4.2 Comparing the Types of Contrastiveness

As has been mentioned before in Section 2.1, the concluding findings in literature on which the formalisations are based on are similar to the findings in [BB22b]. In this section, the specifics of the similarities and differences between the types of contrast found in this research and in [BB22a] will be discussed.

The main difference over all contrasts, as has been mentioned in the previous section, is that my work does not need an extension to be specified. My contrasts compare all sufficient admissible extensions and give all possible explanations - this is an inherent difference to all contrasts compared to the work in [BB22b], where only one explanation is given in a specified extension.

The first contrast is the nonoccurrence-contrast, which can be compared to Borg and Bex' negation-contrast. Borg and Bex defined this contrast by stating that the acceptance of the argument is compared with its non-acceptance. While I can agree with this definition, mine is slightly more simplified. The sufficient arguments for P make it so that P must be accepted. This is in the definition of sufficiency. Due to the manner in which sufficiency is defined, it also accounts for just the relevant arguments. Having found the sufficient arguments for P, it is only logical that it would follow that the negation of P cannot be accepted. To specify the contrast, one could take the set difference over the sufficient arguments, just like in Borg and Bex' work, but it would make no difference - there can be no overlap between sufficient arguments for P and sufficient arguments for $\neg P$. The contrast within the question (comparing P against its negation) is implied due to the sufficiency.

The second contrast is the property-contrast, to be compared against Borg and Bex' despite and within-contrast. They defined this contrast as the reasons for acceptance of the fact that are not part of the acceptance explanation of the foil. They take the set difference, and thus the explanation gives the arguments for why the fact is accepted, which do not include arguments for why the foil could be accepted. This definition I find myself to agree with. To explain a contrastive question, the resulting explanation should be a subset of the possible reasons of *"Why argument P?"*, as the contrast makes the question more specific. The way to narrow down this answer is to first list the reasons for argument P, while excluding any arguments that can also hold when the foil, Q, is accepted. What is left are truly the only arguments that hold specifically when argument P is accepted and argument Q is not accepted. In my case, this set difference is iterated over all the explanations for *"Why argument P?"*, resulting in a set of possible explanations.

The third contrast is the object-contrast, which can be compared to Borg and Bex' between-contrast. Borg and Bex make a comparison between two different extensions, explicitly stating the two extensions to be compared. They define this contrast as the difference of the explanation for the fact in one extension and the extension of the foil. I assumed that for this contrast fact and foil must both be accepted, together, in at least one admissible extension, otherwise this contrast would be what I defined as a property-contrast. Once again, I find Borg and Bex' definition agreeable, and the one I created has a similar foundation. As my definitions do not include specified extensions, the way to compare the different situations in my case is to first find admissible extensions in which fact is true while foil is not, and then take the set difference for each of those extensions with admissible extensions in which fact and foil are both true. This makes it so that the two possible situations are compared against one another, and by taking the set difference one is left with the arguments that make the two situations different.

The final contrast is the over time-contrast, which is not stated at all in Borg and Bex' research. I can understand why this is, the framework needed to model the over time-contrast would need to be entirely different compared to the other contrasts. This was mentioned in Section 3.3; something akin to incomplete argumentation frameworks would be needed to model time in argumentation. Nonetheless, it was a contrast that was mentioned in various research, so it would be remiss not to address its existence, even if it is outside the scope of this research.

This concludes the section on abstract argumentation. In this section, three of the four contrasts (the nonoccurence-, the property- and the object-contrast) have been modelled in Dung's abstract argumentation framework. The over time-contrast has been shown to not be suitable for modelling in Dung's abstract argumentation frameworks. Afterwards, the contrasts have been compared to previous related research, and it has been shown what new findings my definitions contribute to the field of contrastiveness within abstract argumentation.

4 Contrastiveness in Preference-Based Argumentation

The abstract framework for argumentation introduced by Dung in [Dun95] has led to important advances in the study of argumentation. When there is a need to represent more specific argumentation problems, the fully abstract nature of abstract argumentation may present some hindrances. On that account, it is necessary to look into additions to the framework. As a result, as has been briefly mentioned in Section 3.2, there exist more approaches within abstract argumentation which add to Dung's frameworks. One of those additions, and the one this section focuses on, is that of preferences.

Opinions on preferences in XAI are mixed. Miller stated that preferences probably do not matter [Mil19], and Cyras et al. hardly included preferences in their survey of XAI [CRA$^+$21]. When it comes to argumentation, however, it has been shown in past research ([Bre94], [Cay95]) that preference relations are suitable for comparing arguments, as they allow for more sophisticated and more appropriate handling of conflict resolution under inconsistent beliefs or uncertain knowledge.

Preference-based argumentation is a field within argumentation theory which focuses on incorporating preferences into the process of constructing and evaluating arguments. The advantage of preference-based argumentation is that it recognizes that individuals often have different preferences, values, and subjective judgments that influence their acceptance or non-acceptance of arguments. By considering preferences, it becomes possible to find compromises or find alternative solutions that better align with the preferences of multiple parties. Thus, it can be said that preferences are well-suited for argumentation, and are worthwhile to study. This section will elaborate on the addition of preferences in abstract argumentation and then continue to formalise contrastiveness in a preference-based argumentation setting.

4.1 Preference-Based Argumentation Preliminaries

Acceptability of an argument has so far been based on the existence of (in)direct attackers. For this section, preference orderings will be taken into account for comparing the acceptance arguments. First, however, the preliminaries of preference-based argumentation will be discussed.

Preference relations are capable of changing the acceptability of the arguments in a crucial way. The use of preferences enables modelling a notion of individual defense [AC98]. The main idea of a preference-based argumentation framework is that it accepts an argument if it is not defeated, if it is defended by other arguments, or if it defends itself against all attacks by being preferred to its attackers.

Definition 4.1. A Preference-Based Argumentation Framework (PAF) is a triplet $\langle Args, Att, Pref\rangle$, where $Pref$ is a partial preordering on $Args \times Args$, denoted by $<$.

This definition adds preferences to the already familiar abstract argumentation framework. It does this by defining a preference relation between two arguments: argument A is preferred over argument B is defined if and only if $B < A$. While some previous research on preferences also add various types of attack ([AC97] [AC98]), in this research the idea of preferences will be kept more simplified and the focus will be on adding the preference relations to an abstract argumentation framework. This is done to keep the explanations as general as possible, as to make them more widely applicable.

Abstract argumentation frameworks in this work so far have only dealt with either attacking or defending. By adding preference relations, it is important to make a distinction between an attack or defence, and a *successful* attack or defence.

Definition 4.2. Let $\mathcal{AF} = \langle Args, Att, Pref\rangle$ be an abstract argumentation framework with $<$ a preference relation over $Args$, suppose that A and $B \in Args$. Then:

An attack is considered successful iff A attacks B and there exists no preference relation $A < B$.

As stated in Definition 4.2, there can exist an attack relation between two arguments, for example A attacks B, the attack will only be successful if B is not preferred to A. If B is preferred to A, the attack will not be successful, and A does not defeat B. This general idea of adding preferences will change the composition of admissible extensions. With the addition of preference relations and successful attack and defence, if there exists an attack relation (A, B), and B is preferred to A, the set $\{A, B\}$ will still be conflict-free and thus an admissible extension. This will inherently change what arguments will be part of an explanation, depending on what the preference relations are.

Example 12. *Let us look at the initial example framework in Figure 1. To turn this into a PAF, preference relations can be added. For example, let us add the preference relation $d < a$, which means a is preferred to d. The admissible extensions that remain are $\{\}$, $\{a\}$ and $\{a, c\}$. The attack from argument a is now unsuccessful, as a is the preferred argument, thus $\{a\}$ on its own is an admissible extension, and since d can no longer defeat a, b is undefended and any extension with argument b can no longer be admissible.*

4.2 Explanations in Preference-Based Argumentation

Now that the basic notions of PAFs have been elaborated upon, explanations with preferences can be discussed. A sufficient acceptance explanation for an argument A should be an admissible extension of which all the arguments (indirectly) defend A, while keeping in mind the notion of successful attacks. For PAFs, explanations should not just provide the sufficient arguments, but should also include the relevant preference relations: those that have had influence on the (non-)acceptance of the argument. In order to add these relevant preference relations between arguments into the explanations, the definition of sufficient explanations should be extended to include them. This has been done before by Borg and Bex in [BB]. They extended their explanations such that they become pairs: the first element contains the set of arguments according to their definition of an explanation, and the second element contains the relevant preference relations. My definition is based on their previous work:

Definition 4.3. Let $\mathcal{AF} = \langle Args, Att, Pref \rangle$ be an abstract argumentation framework with $Pref$ a preference relation over $Args$ denoted by $<$, suppose that $P, Q, B, C \in Args$. First, suppose that P is accepted, then:

$$\mathsf{PrefIn}(P) = \{(S, \Theta) \mid S \in \mathsf{SuffIn}(P),\ (C < B) \in \Theta \text{ iff } B \in S\}.$$

Now suppose that Q is not accepted, then:

$$\mathsf{PrefOut}(Q) = \{(S, \Theta) \mid S \in \mathsf{SuffOut}(Q),\ (C < B) \in \Theta \text{ iff } C = Q,\ (B, C) \in Att \text{ and } (C < B)\}.$$

To elaborate on this definition, $\mathsf{PrefIn}(P)$ and $\mathsf{PrefOut}(P)$ return a set of pairs of which the first element of each pair contains the set of arguments as in Definition 3.7, which is called the 'basic explanation', and the second element of each pair contains the relevant preference relations. For an acceptance explanation $\mathsf{PrefIn}(P)$, the relevant preference relations are a set of pairs of arguments, such that the preferred argument is part of the basic explanation. This differs from the definition found in [BB], as my definitions make use of admissible extensions and the argument that is to be explained is always included in the explanation itself. For a non-acceptance explanation $\mathsf{PrefOut}(P)$, the relevant preference relations are a set of pairs of arguments, such that the the non-preferred argument is part of the basic explanation and the attack from the preferred argument is successful due to the preference relation. Aside from simplifying the acceptance explanation and adding a non-acceptance explanation, there are two additional changes in this definition compared to the one it is based on in [BB]. The first is that only sufficient explanations are considered, as they have been defined in Definition 3.7. The second change is that this definition can give various possible explanations; an extension is not specified and the relevant preference relations are found for each possible explanation for argument P. Thus, this definition gives back a set of pairs: one for each explanation rather than just one pair.

4.3 Contrastiveness in Preference-Based Argumentation Frameworks

In this section I will discuss how to apply preference orderings on the previously established definitions of contrastiveness and see whether the explanations still perform well. By this, I mean that the explanations effectively achieve their intended purpose: they find the sufficient arguments that explain the contrast, and find any preference relations that have had influence on that contrast. In this section, one by one the three types of contrastiveness (nonoccurence, property and object) will be discussed to see how the addition of preferences influences the contrastive explanations.

The first contrast to discuss will be the nonoccurence-contrast from Definition 4.4. The goal here is to find the set of pairs, where the first part of each pair is a sufficient admissible extension as explanation, and the second part is the relevant preference relations for that explanation. As it suffices to find the sufficient acceptance explanation for a nonoccurence-contrast, the definition for this contrast will be the same as Definition 4.3.

Definition 4.4. Let $\mathcal{AF} = \langle Args, Att, Pref\rangle$ be an abstract argumentation framework with *Pref* a preference relation over *Args* denoted by $<$, suppose that $P \in Args$ Then:

$$\mathsf{PrefNonCont}(P) = \mathsf{PrefIn}(P).$$

Example 13. *Consider the abstract argumentation framework in Figure 4, and add the preference relations $Pref = \{(d < a), (e < f)\}$. Asking the question "Why a rather than not a?" will result in the following explanation. First, one finds the basic explanations; the set of all possible sufficient explanations: which in this case is the set of sets $\{\{a\}, \{a, c\}\}$. Then, one finds the second part of the pair; the relevant preference relations for each explanation. In the case of the preference relation $(d < a)$, the preferred argument is in both of the explanations. Thus, this preference relation is relevant for both possible explanations. This is not the case for the preference relation $(e < f)$. As a result, the explanation for the question "Why a rather than not a?" will result in the set of pairs $\{(\{a\}, \{(d < a)\}),\ (\{a, c\}, \{(d < a)\})\}$ as possible explanations.*

As can be gathered from Example 13, the definition for the nonoccurence-contrast still performs well - even when adjusting to PAFs. The next contrast to discuss is the property-contrast from Definition 3.9. For all possible sufficient acceptance explanations for the fact, the difference should be taken with the union of all possible acceptance explanations of the foil. However, now the relevant preferences should be included in the explanation. Thus, while applying the definition for property-contrast one should find the relevant preference relations based on Definition 4.3. This results in the following definition:

Definition 4.5. Let $\mathcal{AF} = \langle Args, Att, Pref\rangle$ be an abstract argumentation framework with *Pref* a preference relation over *Args* denoted by $<$, suppose that $P, Q, B, C \in Args$. Then:

$$\begin{aligned}\mathsf{PrefPropCont}(P, Q) = \{(S, \Theta) \mid S \in \mathsf{PropCont}(P, Q),\ (C < B) \in \Theta \text{ iff } B \in S \\ \text{or } (C = Q,\ (B, C) \in Att \text{ and } (C < B))\}.\end{aligned}$$

Compared to the previous definition of which preferences should be included in the explanation, one addition has been made. Aside from including the preferences of which the preferred argument is in the basic explanation, one should also include the arguments which are preferred to the foil and as a result, successfully attack the foil. In this manner, one will receive, per explanation, the preferences that have influence on the acceptance of the foil and the non-acceptance of the fact.

Example 14. *Consider the abstract argumentation framework in Figure 4, and add the preference relations $Pref = \{(e < f), (e < c)\}$. Asking the question "Why c rather than e?" will result in the following explanation. First, one finds the contrastive explanations according to Definition 3.9, which results in the possible explanations $\{\{a, c\},\ \{a, c, f\}\}$. Now, for each possible explanation, the relevant preference relations will be added to create a set of pairs. For a preference relation to be in Θ, the preferred argument of that preference relation should be in the basic explanation. For the preference relation $(e < c)$, this is the case for both of the explanations. For the preference relation $(e < f)$, this is only the case for the explanation $\{a, c, f\}$. However, an additional possibility for a preference relation to be considered relevant, is if the second argument is the foil, there exists an attack relation between the first and second argument and this attack relation succeeds due to the preference relation. This is the case for both the preference relations. Combining these findings results in the explanations $\{\{a, c\}, \{(e < f), (e < c)\}),\ (\{a, c, f\}, \{(e < f), (e < c)\})\}$.*

Example 14 shows that the relevant preference relations for a property-contrast can be found for each explanation. This example shows that preference relations that influence the non-acceptance of the foil are accounted for as well, and it can handle indirect attacks influenced by preferences, which still are relevant. The definition for the property-contrast thus still performs well, even when adding preference relations. This only leaves us with the last contrast to discuss in this setting, which is the object-contrast as defined in Definition 3.10. This one will be handled in quite a similar way as the previous contrast: after having found the explanations according to Definition 3.10, for each explanation the relevant preference relations will be found, resulting in a set of pairs of explanations and preference relations. This results in the following definition:

Definition 4.6. Let $\mathcal{AF} = \langle Args, Att, Pref\rangle$ be an abstract argumentation framework with *Pref* a preference relation over *Args* denoted by $<$, suppose that $P, Q, B, C \in Args$. Then:

$$\mathsf{PrefObjCont}(P, Q) = \{(S, \Theta) \mid S \in \mathsf{ObjCont}(P, Q),\ (B, C) \in \Theta \text{ iff } (C = Q,\ (B, C) \in Att \text{ and } (C < B))\}.$$

This definition once again has been adjusted from the original to find the preference relations that are relevant to the current contrast, which is the object-contrast. Now, for a preference to be considered relevant to the explanation, the non-preferred argument of the preference should be the foil, the preferred argument should attack the non-preffered argument and that attack should succeed due to the preference relation. Unlike the previous contrast, preferences related to the fact are not necessarily considered relevant. In the object-contrast, two situations are compared - one where fact is true and foil is not, and one where fact and foil both are true. Thus, reasons for why the fact is accepted are not as relevant, since it is accepted in both situations, and one wants to find the differences between the two situations as an explanation for why the foil is accepted in one but not the other. Thus, relevancy of preferences relies more on the foil in this situation.

Example 15. *Consider yet again the abstract argumentation framework in Figure 4, and add the preference relations* $Pref = \{(a < b), (c < e)\}$*. The question to explain is "Why is* b *true and not* f*, rather than* b *and* f *being true?". The sufficient contrastive explanation for* b *and not* f *would be the set* $\{b, d, e\}$*, and the set difference should be taken with the sufficient admissible extension in which* b *and* f *are both true, which is the set* $\{b, d, f\}$*. This leaves us with* $\{e\}$ *as the sufficient contrastive explanation. The next step is to find the relevant preference relations. For a preference relation to be in* Θ*, the non-preferred argument should be the foil, should be attacked by the preferred argument and that attack succeeds due to the preference relation. This is not the case for either preference relation. This means that for this explanation, there are no relevant preference relations, thus the explanation would be* $(\{e\}, \emptyset)$*.*

Example 15 shows that the definition of the object-contrast performs well and also accounts for when there are no relevant preferences - making sure that those are not included in the contrastive explanation. This definition only includes preferences that are relevant to the contrast to be explained.

Example 16. *A second example will be based on the abstract argumentation framework in Figure 2, now with the preference relations* $\{(b < a), (c < d)\}$*. The question to explain is "Why is* a *true and not* c*, rather than* a *and* c *being true?". The contrastive explanation for* a *and not* c *would be the set* $\{a, d, e\}$*, and the set difference should be taken with the sufficient admissible extension in which* a *and* c *are both true, which is due to the added preference relations the empty set* $\{\emptyset\}$*. This leaves us with* $\{a, d, e\}$ *as the sufficient contrastive explanation. Now to find the relevant preference relations, the non-preferred argument should be the foil, should be attacked by the preferred argument and that attack succeeds due to the preference relation. This is only the case for the preference relation* $(c < d)$*, thus it will be added to the explanation. As a result, the explanation for this question will be* $\{(\{a, d, e\}, \{(c < d)\})\}$*.*

This concludes the modelling of contrastiveness in preference-based argumentation frameworks. At the hand of the new definitions accounting for relevant preference relations (based on previous work in [BB]), I showed that my definitions for the different contrasts perform well, and can provide explanations with relevant preferences.

4.4 Evaluation of Contrastiveness in Preference-Based Argumentation

In the previous sections, definitions for various types of contrastiveness have been constructed and applied in preference-based abstract argumentation frameworks. So far, these definitions seem to perform well in the various examples that have been provided: they were shown to be capable of giving suitable contrastive explanations.

The findings on how to model contrastiveness in abstract argumentation could be compared to previous work (as was done in Section 3.4), as that research was not the first of its kind. However, contrastiveness based on literature has so far not been studied in the setting of preference-based abstract argumentation frameworks. This is a new field of research and thus a comparison between previous work is not viable. However, it is still vital to evaluate this work in some manner. Due to the scope of this thesis, an elaborate survey is unfortunately not feasible. Instead, in this section, various real-world problems will be cited and the definitions of contrastiveness with preferences will be employed to see whether they will come up with satisfactory explanations to the various situations.

In AI, the proper modelling of argumentation for decision-making is an important current issue, and given the importance of purpose and value in the law, the law provides an excellent testing ground for the explanations based on argumentation. Previously, formal argumentation for legal practical reasoning has been studied ([BP10], [BCPV11], [Wal09]). For this reason and for the sake of evaluating my definitions of contrastiveness in preference-based argumentation, legal practical reasoning will be used as the grounds for the example situations. Example cases will be taken from [BPS09], [Pra15] and [BB21b].

The first example will be based on a widely cited case, the one of *Keeble v Hickeringill* (1707). It is a well-known case from Anglo-American property law on ownership of wild animals. To describe the case shortly, a pond owner (who is the plaintiff) placed a duck decoy in his pond with the intention to sell the caught ducks for a living. However, the defendant used a gun to scare away the ducks, purely to interfere with the plaintiff's business. The plaintiff in this case argued that people should be protected when they are pursuing their livelihood, and that he was hunting on his own land. The defendant claimed that the ducks were yet to be caught, so they did not belong to anyone yet. A possible abstract argumentation framework of this case can be seen in Figure 5. The arguments displayed are as follows:

$P1$: Plaintiff wants to sell caught ducks in owned pond for a living.

$D1$: Defendant wants to stop this, and scares ducks away by using a gun.

$P2$: People should be protected when pursuing their livelihood.

$P3$: Plaintiff was hunting on his own land.

$D2$: Plaintiff had no rights to the ducks, as they were not yet caught.

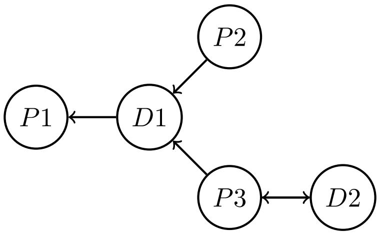

Figure 5: an example abstract argumentation framework of the case Keeble v Hickeringill (1707)

At the hand of this case, it is possible to try and evaluate how the definitions of contrastiveness for PAFs perform in a scenario that is based on real events.

Example 17. *Consider the scenario displayed in Figure 5 based on the case Keeble v. Hickeringill. The case in question was ruled in favour of the plaintiff by the courts. From this conclusion, we can suppose that there existed a preference relation $D2 < P3$; argument $P3$ was preferred to $D2$. One could ask for an explanation in response to the result of the court ruling: "Why could the plaintiff sell caught ducks, but could the defendant not scare them away?". In terms of the AF, this question would "Why $P1$, but not $D1$?". In this case, the type of contrast is a property-contrast. Given the preference relation $D2 < P3$, $\mathsf{SuffIn}(P1)$ gives the sets $\{\{P1, P2, P3\}, \{P1, P2\}, \{P1, P3\}\}$, and $\mathsf{SuffIn}(D1)$ gives the set $\{\emptyset\}$. Taking the set difference between the two results in the sets $\{\{P1, P2, P3\}, \{P1, P2\}, \{P1, P3\}\}$. For a preference relation to be considered relevant, one possibility is that the preferred argument of that preference relation should be in the basic explanation. This is the case for preference relation $D2 < P3$ for two of the possible explanations, thus the explanation results in the following pairs: $\{(\{P1, P2, P3\}, \{(D2 < P3)\}),\ (\{P1, P2\}, \{\emptyset\}),\ (\{P1, P3\}, \{(D2 < P3)\})\}$.*

Example 17 shows a contrastive question and suitable explanations in a situation based on a real-world example. At the hand of this example, it has been shown that contrastive explanations can be given, while including only the preferences that are truly relevant to the contrast and the explanation. In comparison to contrastive explanations in regular abstract argumentation, now additional information is included: preferences that have influence on the acceptance of the fact or the non-acceptance of the foil. However, while this example does show that my definitions for contrastiveness in preference-based argumentation frameworks perform well even in a situation based on real-world cases, it does not necessarily show how preferences benefit explanations - and how vital it is for them to be included. This will be illustrated more thoroughly in the next examples.

Another well-known case is the one of *Olga Monge v Beebe Rubber Company* (1974). Olga Monge was employed at will by Beebe Rubbber Company, meaning that she was employed indefinitely. Given the common law rule at the time, every indefinite employment contract is terminable at will by either party. At some point, Olga Monge was fired for no given reason by her foreman, and Olga claimed that it was due to her refusing to go out with him. If that was the reason, it would have been a breach of contract. The common law rule does not apply if the employee was fired in bad faith, malice or retaliation. The court accepted this as the reason for being fired, and had to decide whether to follow the old rule and state that there was no breach of contract, or distinguish the rule into a new rule by adding an exception in case the employee was fired in bad faith, malice or retaliation.

The case of Olga Monge would consist of the following arguments in an abstract argumentation setting:

A: The new rule should be adopted as the valid rule.

B: Someone employed at will can be fired for any reason or no reason at all unless the employee is fired in bad faith, malice, or retaliation.

C: Olga Monge could not be fired for no reason.

D: The old rule should be adopted as the valid rule.

E: Someone employed at will can be fired for any reason or no reason at all.

F: Olga Monge could be fired for no reason.

A graphical representation of these arguments can be found in Figure 6.

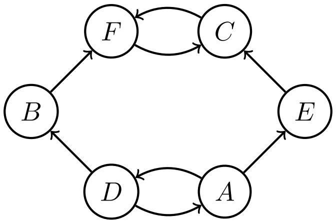

Figure 6: an example abstract argumentation framework of the case Monge v Beebe Rubber Company

At the hand of this situation, it is possible to once again evaluate possible contrastive explanations in this setting. At the time, the court ruled in favour of Monge, and thus an exception to the common law was added. From this, we can assume that there was a preference relation $A < D$, where it was preferred to adopt the new rule as the valid rule rather than the old rule.

Example 18. *Consider the scenario displayed in Figure 6, and consider the preference relation* $(D < A)$*, where argument* A *is preferred to argument* D*. One can ask the contrastive question "Why could Olga Monge not be fired for no reason, rather than it being possible?". In terms of the AF displayed, this question concerns argument* C*: "Why* C*, rather than* $\neg C$*?" This contrastive question consists of a nonoccurence-contrast.* $\mathsf{SuffIn}(C)$ *gives the sets* $\{\{A, C\}, \{A, B, C\}\}$*. For a preference relation to be considered relevant, the preferred argument of that preference relation should be in the basic explanation. This is the case for preference relation* $(D < A))$ *for both possible explanations. Thus, the resulting possible explanations are* $\{(\{A, C\}, \{(D < A)\}), (\{A, B, C\}, \{(D < A)\})\}$*.*

Example 18 shows the importance of preferences in a real-world situation. By just looking at the framework pictured in Figure 6, one outcome is not stronger than the other. Weighing arguments F and C against one another, they come out evenly strong. The arguments *Olga Monge could be fired for no reason* and *Olga Monge could not be fired for no reason* are (indirectly) defeated by and defeat each other - so how could one conclude one argument over the other? The only manner to achieve this is by including preferences, as has been done at the time. The court ruled in favour of Monge not being able to be fired for no reason, and decided that the new rule should be adopted as the valid rule, thus there was a preference present. The only way this case and its ruled outcome could possibly be modelled in an abstract argumentation setting, is by including preferences. An explanation for the argument *Olga Monge could not be fired for no reason* should thus include a preference relation, as that is what made the decision decisive for the court at the time.

The final example situation to evaluate contrastiveness in PAFs will be the one sketched in [BB21b]. Compared to the earlier evaluations, this is a more elaborate example. This example considers malafide webshops and is build on the following arguments:

$A1$: A complaint was filed.

$A2$: The complaint is retracted.

$A3$: The url is suspicious.

$A4$: The address is registered at the chamber of commerce.

$A5$: The webshop owner is known by the police.

$A6$: The registration was recently retracted.

$B1$: A complaint was filed, thus an investigation is done.

$B2$: The complaint was retracted, an investigation is not done.

$B3$: The webshop owner is known by the police, thus the complaint can not be retracted.

$B4$: An investigation is done and the url is suspicious, thus the webshop is malafide.

$B5$: The address is registered at the chamber of commerce, thus the webshop is not malafide.

$B6$: The registration was recently retracted, thus the address is not registered at the chamber of commerce.

A graphical representation of the corresponding argumentation framework can be found in Figure 7, as it was depicted in [BB21b]. The final example to be discussed will regard this framework.

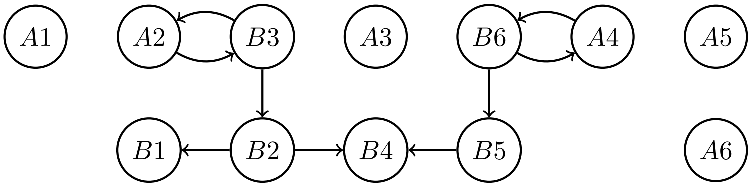

Figure 7: an abstract argumentation of the malafide webshop case based on the example provided in [BB21b]

Example 19. *Consider the scenario displayed in Figure 7 based on malafide webshop cases at the Netherlands Police. Consider adding the preference relation* $(A2 < B3)$*, where argument* $B3$ *is preferred to* $A2$*. The Netherlands Police could prefer not retracting an already made complaint when the webshop owner is already known by the police, as an extra measure of safety.*

One can ask the contrastive question "Why is the workshop found to be malafide, rather than not having done an investigation?". In terms of the AF displayed, this question concerns arguments $B4$ *and* $B2$*: "Why* $B4$ *rather than* $B2$*?". The type of contrast to consider here is a property-contrast. The sufficient acceptance explanation for* $B4$ *would be the set* $\{B3, B4, B6\}$*, and the set difference should be taken with the sufficient explanation for the acceptance of* $B2$*, which is, due to the added preference relation, the empty set* $\{\emptyset\}$*. This leaves us with* $\{B3, B4, B6\}$ *as the sufficient contrastive explanation. Now it is important to find the relevant preference relations. For a preference relation to be in* Θ*, the preferred argument should be in the basic explanation, or the non-preferred argument should be the foil, should be attacked by the preferred argument and that attack succeeds due to the preference relation. This is indeed the case for the given preference relation. This means that for this explanation, the explanation would be* $(\{B3, B4, B6\}, \{(A2 < B3)\})$*.*

Example 19 shows how one could consider what kind of approach one prefers to take in a decision-making scenario. This makes preference-based argumentation much more flexible than regular abstract argumentation: within the same framework, different preferences can model different choices and approaches. This example considers a lower-risk approach: if a complaint is made and the address is known by the police, it is preferred to *not* be able to retract that complaint. As was stated in the beginning of Section 4, by considering preferences, it becomes possible to find compromises or find solutions that better align with the preferences of multiple parties. Especially when it comes to topics as fraud or malafide webshops, there are various reasons as why one would want to take a more careful approach and prefer investigating more over not investigating. Preference-based argumentation can model such choices better than regular abstract argumentation can. If such choices are made, however, they should also be included in the explanations, otherwise that explanation would not reflect the full situation.

At the hand of Examples 17 to 19, it has been shown that even in situations based on real-world situations, my definitions of contrastiveness for preference-based argumentation still perform well, as they are capable of giving suitable explanations while including the relevant preferences. All these examples combined have discussed all the three contrasts (nonoccurence-, property- and object-contrast), as well as situations where there are relevant preferences and situations where there are not. From this, one can conclude that the contrasts have been evaluated in preference-based argumentation. As has been stated in Section 4.1, preference relations are relevant for comparing arguments for multiple reasons. Including preferences in explanations can help make abstract arguments more understandable to human users, as they are capable of providing a context for arguments, and allow for more subjective judgements to be made. Preferences also allow for more flexibility in abstract argumentation, as it allows for different interpretations and different possible explanations. As has been shown in various examples in the previous sections, preferences can strongly influence the composition of admissible extensions, which explanations are based on in this research. This flexibility allows for a more nuanced approach to explanation, and including preferences in explanations can allow for a more adaptable and customisable reasoning process, where users can adjust the evaluation of arguments based on their preferences. This also applies to real-world decision-making scenarios, as preferences are inherent to many of them. As has been stated previously, this is especially well-illustrated in example 18: without the inclusion of preferences, the court would not have been able to properly explain why one solution was chosen over the other. The preference made the difference in the outcome, and an abstract argumentation framework without preferences could thus not fully explain why the choice has been made. In the real world, individuals often have preferences and biases, and including preferences in explanations can help capture this complexity by accounting for the subjective nature of human decision-making. By incorporating preferences, abstract argumentation can better reflect the diverse and dynamic nature of real-world decision-making, and thus explain such situations better.

At the hand of the evaluations in this section, I have shown that preferences are worthwhile to study for contrastive explanations, as they allow for more sophisticated and more appropriate handling of conflict resolution under inconsistent beliefs or uncertain knowledge [AC02]. To the best of my knowledge, contrastiveness based on literature in the social sciences and humanities has not yet been studied within preference-based argumentation frameworks. But with this evaluation, it has been shown that my definitions for contrastive explanations can still account for abstract argumentation frameworks that consider preference relations, and that preferences can better reflect certain choices made in specific situations, which regular abstract argumentation cannot always account for. In other words, preferences can provide additional information to better reflect the actual situation, and they allow for a more flexible modelling of argumentation. Thus, when considering contrastive explanations in argumentation, preferences should be accounted for as well.

5 Contrastiveness ASPIC$^+$

In the previous sections, all examples of argumentation have been within the field of abstract argumentation: the arguments within the framework are abstract entities. Abstract argumentation is indispensable for argumentation theories, as has been stated in previous sections, but it provides no information on the structure of the arguments, and thus gives no guidelines on how to model actual argumentation problems [MP14b]. How can abstract argumentation model arguments that are mutually consistent, or how can it model deductive inference? Abstract argumentation is very versatile and widely applicable due to its abstract nature, but it does suffer from various shortcomings because of that as well.

Structured argumentation, as the name suggests, steps away from the fully abstract nature of abstract argumentation, and its goal is to provide some guidance on the structure of arguments. It does this by including internal structures of an argument [BGH$^+$14]. Structured argumentation offers a more detailed formalisation of arguments than abstract argumentation. Knowledge is represented in a formal language, and the premises, conclusions and relationships between premise and conclusion of arguments are made explicit. Overall, structured argumentation provides a systematic framework for modeling, analyzing, and evaluating arguments, which promotes clearer reasoning, more informed discussions, and a better understanding of complex issues.

In the next section, the structured argumentation framework ASPIC$^+$, as defined in [MP13], will be discussed. ASPIC$^+$ was chosen as the structured argumentation framework to use, as it is a widely used and cited structured argumentation framework, and applications based on ASPIC$^+$ are used in the real world [OBBT22].

5.1 ASPIC$^+$ Preliminaries

ASPIC$^+$ almost fully abstracts from the nature of the logical language and the inference rules and can be instantiated in various ways [BGH$^+$14]. As a short history on the origin of ASPIC$^+$, Lin and Shoham were the first to propose the idea of abstraction in structured argumentation. They expanded upon the idea of abstract argumentation structures with strict and defeasible rules and showed how previously established nonmonotonic logics could be reconstructed as such structures [LS89]. Vreeswijk further developed these ideas into his abstract argumentation systems. His ideas are included in today's ASPIC$^+$ framework [Pra18], which is the one used in this research.

The ASPIC$^+$ framework was designed to compromise between Dung's abstract AFs and abstraction between concrete logics [MP14b]. The ASPIC$^+$ framework is based on two main ideas: the first idea is that conflicts between arguments are often seen to be resolved with explicit preferences, and the second idea is that arguments should be built with two kinds of inference rules: strict rules, where the premises explicitly guarantee the conclusion, and defeasible rules, where the premises only provide an option for the conclusion [MP14b].

ASPIC$^+$ offers an advanced and flexible framework for structured argumentation with several benefits. It offers expressiveness, flexibility, and decision support, and as a result it finds applications in domains such as law and policy-making ([OBBT22], [Pra15]). ASPIC$^+$ supports decision-making by providing a structured approach to analyze and evaluate arguments while considering preferences and constraints. It can help in understanding the underlying reasoning, assumptions, and evidence behind arguments, and aid in explanation processes in decision-making contexts. Additionally, ASPIC$^+$ can model complex scenarios with incomplete and uncertain information, while incorporating preferences and constraints. It is adaptable to various argumentation contexts and domains, making it versatile and thus suitable for a wide range of applications.

Due to these reasons and the existence of real-world applications based off of ASPIC$^+$, ASPIC$^+$ proves to be a valuable asset in the study of XAI. Therefore, the topic of contrastiveness should also be studied within an ASPIC$^+$ setting. Contrastiveness itself has been studied within ASPIC$^+$, namely in [BB21b]. However, as Borg and Bex stated, much more can be said about the individual concepts of contrastiveness than was presented in their research. To the best of my knowledge, contrastiveness based on findings in the social sciences and humanities, with varying types of contrast, has not yet been studied within ASPIC$^+$, and as a result, the findings of this section will make a valuable contribution to the field of XAI.

In this section, the most important aspects of ASPIC$^+$ will be given as they are defined in [MP14b]. To start, we need an argumentation system:

Definition 5.1. An argumentation system is a triple $AS = (\mathcal{L}, \mathcal{R}, n)$, where:

1. $\mathcal{L}$ is a logical language closed under classical negation ($\neg$).
2. $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible inference rules ($\mathcal{R}_d$) of the form $\phi_1, ..., \phi_n \rightarrow \phi$ and $\phi_1, ..., \phi_n \implies \phi$ respectively (where $\phi_1, ..., \phi_n, \phi \in \text{wff}(\mathcal{L})$ and $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$.
3. n is a partial function such that $n : \mathcal{R}_d \longrightarrow \mathcal{L}$.

In addition to the argumentation system, one also needs to specify a knowledge base: a body of information from which the premises of an argument can be taken. This knowledge base distinguishes ordinary premises (which can be attacked) and premises that are axioms (which cannot be attacked). Once again, this definition is defined as it is in [MP14b]:

Definition 5.2. A knowledge base in an $AS = (\mathcal{L}, \mathcal{R}, n)$ is a set $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets $\mathcal{K}_n$ (the axioms) and $\mathcal{K}_p$ (the ordinary premises).

One is free to decide what is an axiom and what is an ordinary premise, and how to specify the strict and defeasible rules. It should be noted that conclusions of arguments in the same extension should be mutually consistent, and they ought to be closed under strict inference rules [MP14b].

With these aspects defined, we can move on to an argumentation theory. An argumentation theory is the combination of an argumentation system with a knowledge base:

Definition 5.3. An argumentation theory is a tuple $AT = (AS, \mathcal{K})$ where AS is an argumenation system and $\mathcal{K}$ is a knowledge base in AS.

Arguments are constructed from an argumentation theory as follows:

Definition 5.4. An argument A on the basis of an argumentation theory with a knowledge base $\mathcal{K}$ and an AS $(\mathcal{L}, \mathcal{R}, n)$ is:

- ϕ if $\phi \in \mathcal{K}$ with: $\mathsf{Prem}(A) = \{\phi\}$, $\mathsf{Conc}(A) = \{\phi\}$, $\mathsf{Sub}(A) = \{\phi\}$, $\mathsf{DefRules}(A) = \{\emptyset\}$, $\mathsf{TopRule}(A)$ = undefined.
- $A_1, ..., A_n \rightarrow \psi$ if $A_1, ..., A_n$ are arguments such that there exists a strict rule

  $\mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n) \rightarrow \psi \in \mathcal{R}_s$.
  $\mathsf{Prem}(A) = \mathsf{Prem}(A_1) \cup ... \cup \mathsf{Prem}(A_n)$.
  $\mathsf{Conc}(A) = \psi$,
  $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup ... \cup \mathsf{Sub}(A_n) \cup \{A\}$.
  $\mathsf{DefRules}(A) = \mathsf{DeRules}(A_1) \cup ... \cup \mathsf{DefRules}(A_n)$,
  $\mathsf{Conc}(A_n) \implies \psi$.

- $A_1, ..., A_n \implies \psi$ if $A_1, ..., A_n$ are arguments such that there exists a defeasible rule $\mathsf{Conc}(A_1)$,..,$\mathsf{Conc}(A_n) \implies \psi$ in $\mathcal{R}_d$.

The above notation can be generalised to sets. For example, where S is a set of arguments $\mathsf{Prem(S)} = \bigcup\{\mathsf{Prem}(A) \mid A \in \mathsf{S}\}$ and $\mathsf{Conc(S)} = \{\mathsf{Conc}(A) \mid A \in \mathsf{S}\}$.

Example 20. *In Figure 8 an example of an argument in ASPIC$^+$ is pictured. This figure considers a knowledge base in an argumentation system with $\mathcal{L}$ consisting of $\{p\}$, with $\mathcal{R}_s = \{s1\}$ and $\mathcal{R}_d = \{d1\}$, where $d1 : \ p \implies q$ and $s1 : \ p, q \rightarrow r$.*

The premise (p) is displayed at the bottom and the conclusions (q and r) are displayed at the top of the argument graph. In actuality, three arguments are displayed here: A_1: p, A_2: $A_1 \implies$ q and A_3: A_1, $A_2 \rightarrow$ r.

As is visible in Figure 8, defeasible rules and ordinary premises are displayed with dashed lines, while strict rules and axioms are displayed with a regular lines.

In addition to the different kinds of inference rules and arguments, there are also different kinds of attacks: undercuts, rebuttals and underminers. An argument is undercut if it is attacked on one of its inference rules, an argument is rebutted if it is attacked on its conclusion and an argument is undermined if it is attacked on its premise. Arguments can only be undercut on the application of defeasible inference rules, but not on strict rules [MP14b].

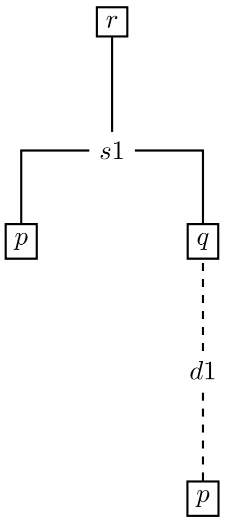

Figure 8: An example argument in ASPIC$^+$

Definition 5.5. A attacks B if A undercuts, rebuts or undermines B, where

- A undercuts argument B (on B') iff $\mathsf{Conc}(A) = \neg n(r)$ for some $B' \in \mathsf{Sub}(B)$ such that B''s top rule r is defeasible.
- A rebuts argument B (on B') iff $\mathsf{Conc}(A) = \neg\phi$ for some $B' \in \mathsf{Sub}(B)$ of the form $B_1, ..., B_n \implies \phi$.
- A undermines argument B (on ϕ) iff $\mathsf{Conc}(A) = \neg\phi$ for an ordinary premise ϕ of B.

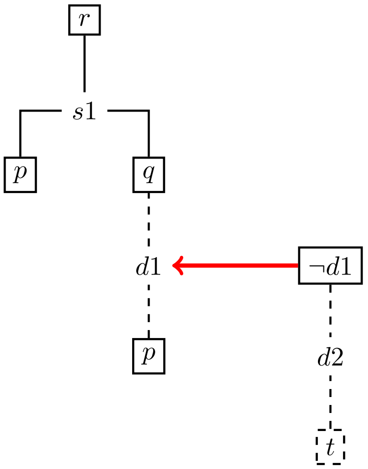

Figure 9: An example undercutting attack in ASPIC$^+$

Example 21. *In Figure 9 one can see an example of multiple arguments and one attack in ASPIC$^+$. This figure considers a knowledge base in an argumentation system with $\mathcal{L}$ consisting of $\{p, t\}$, with $\mathcal{R}_s = \{s1\}$ and $\mathcal{R}_d = \{d1, d2\}$, where $d1: \ p \implies q$, $d2 : t \implies \neg d1$ and $s1: \ p, q \rightarrow r$.*

The premises (p and t) are displayed at the bottom and the conclusions $(r, q, \neg d1)$ are displayed higher up the argument graph. In actuality, four arguments are displayed here: A_1: p, A_2: $A_1 \implies q$, A_3: A_1, $A_2 \rightarrow r$ and A_4: $t \implies \neg d1$.

The type of attack illustrated here is an undercutting attack, as the defeasible conclusion $\neg d1$ attacks the defeasible inference rule $d1$.

Finally, the definition will be given for an argumentation theory of an argumentation framework. ASPIC$^+$ argumentation theories have corresponding Dung-style argumentation frameworks [BB21c]:

Definition 5.6. From an AT the corresponding AF can be derived such that $\mathcal{AF}(AT) = \langle Args, Att\rangle$, where Args is the set of arguments constructed from AT and $(A, B) \in Att$ iff $A, B \in Args$ and A attacks B as defined in Definition 5.5.

With these basic notions of ASPIC$^+$ defined, first, the concept of regular explanations will be defined in Section 5.2, after which in Section 5.3 the definitions of contrastiveness will be modelled in this structured argumentation setting. Finally, in Section 5.4 these contrastive definitions will be evaluated.

5.2 Explanations in ASPIC$^+$

In Section 3.5, the basic definition for an explanation in abstract argumentation was given: in this research, an explanation for the acceptance of an argument is considered all admissible extensions of which it is a part, and a non-acceptance explanation consists of all the admissible extensions of which the argument is not a part and which attack the argument. Later on, sufficiency was an added requirement in Definition 3.7 to narrow down the set of possible explanations to only include sufficient explanations.

For ASPIC$^+$ it is important to once again look at what consists of an explanation, and start from the beginning. First, a notion of acceptance and non-acceptance will be defined, based on [BB21c]:

Definition 5.7. Let $\mathcal{AF}(AT) = \langle Args, Att\rangle$ be an AF, based on AT, let $Adm(\mathcal{AF})$ be the collection of all admissible extensions of $\mathcal{AF}$ and let $\phi \in \mathcal{L}$. Then ϕ is:

- accepted: if $\phi \in \bigcup \mathsf{Conc}(Adm(\mathcal{AF}(AT)))$, that is: there is some argument with conclusion ϕ that is part of an admissible extension;
- non-accepted: if $\phi \notin \bigcap \mathsf{Conc}(Adm(\mathcal{AF}(AT)))$, that is: there is some admissible extension without an argument with conclusion ϕ.

Just as has been done in Sections 3 and 4 about explanations in abstract argumentation, the focus of explanations in ASPIC$^+$ will be based on admissible extensions as well.

Now that acceptance itself has been defined, it is possible to move forward to actual acceptance and non-acceptance explanations for formulas. These explanations are defined in terms of a function $\mathbb{F}$, which determines what elements of the arguments the explanation presents. In ASPIC$^+$, the structure of the arguments is known and can be used in the explanations. By using the function $\mathbb{F}$, explanations can consist of the premises of arguments, the sub-conclusions of arguments or the arguments themselves. The function $\mathbb{F}$ can be instantiated in different ways, as they have been introduced in [BB21a]:

- $\mathbb{F} = \mathsf{id}$, where $\mathsf{id(S)} = \mathsf{S}$. Then explanations are sets of arguments.
- $\mathbb{F} = \mathsf{Prem}$. Then explanations only contain the premises of arguments.
- $\mathbb{F} = \mathsf{SubConc}$, where $\mathsf{SubConc}(A) = \{\mathsf{Conc}(B) \mid B \in \mathsf{Sub}(A), \mathsf{Conc}(B) \notin \mathcal{K} \cup \{\mathsf{Conc}(A)\}\}$. Then the explanation contains the sub-conclusions that were derived in the construction of the argument.

Formula explanations for ASPIC$^+$ differ in two ways from the explanations in Section 3 and Section 4: first, they involve the application of the function $\mathbb{F}$ (which can only be applied to formula explanations), and second, they require consideration of the arguments for ϕ. Acceptance and non-acceptance explanations are defined as the following:

Definition 5.8. Let $\mathcal{AF}(AT) = \langle Args, Att\rangle$ be an AF, based on AT and suppose that $\phi \in \mathcal{L}$ is accepted. Then:

$$\mathsf{In}(\phi) = \{\mathbb{F}(S) \mid S \text{ is an admissible extension, } \exists A \in S, \text{ and } \mathsf{Conc}(A) = \phi\}.$$

Now let $\mathcal{AF} = \langle Args, Att\rangle$ be an AF, based on AT and suppose that $\phi \in \mathcal{L}$ is not accepted. Then:

$$\mathsf{Out}(\phi) = \{\mathbb{F}(S) \mid S \text{ is an admissible extension, } \exists B \in Args,\ \exists A \in S,\ A \times B \in Att \text{ and } \mathsf{Conc}(B) = \phi\}.$$

An acceptance explanation for a formula contains all the admissible extensions of which the formula is the conclusion of at least one of the arguments. A non-acceptance explanations for a formula contains all the admissible extensions of which the formula is not a part, which also attack the arguments with ϕ as its conclusion. The function $\mathbb{F}$ can be applied to the explanations.

Note that formula $\mathbb{F}$ can also be applied to a collection of sets. To apply $\mathbb{F}$ to a collection of sets, the following definition will be used:

Definition 5.9. Suppose $\mathcal{S}$ is a collection of sets. Then:

$$\mathbb{F}(\mathcal{S}) = \{\forall S \in \mathcal{S} \mid \mathbb{F}(S)\}.$$

In other words, if function $\mathbb{F}$ is applied to a collection of sets, it means that it will be applied to every set within that collection.

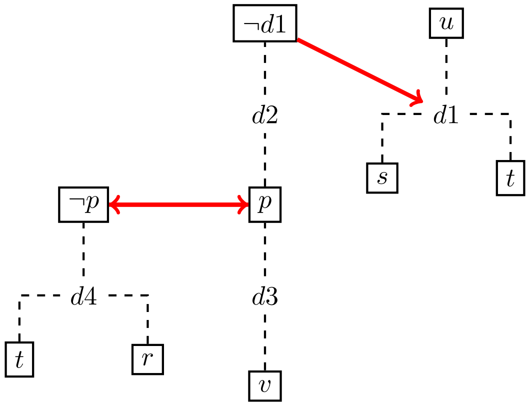

Figure 10: Another example argumentation system in ASPIC$^+$

Example 22. *Let $AS = \langle \mathcal{L}, \mathcal{R}, n \rangle$ where the rules in $\mathcal{R}$ are such that, with $\mathcal{K} = \mathcal{K}_n = \{r, s, t, v\}$ and with $\mathcal{R}_d = \{d_1, d_2, d_3, d_4\}$, the following arguments can be derived:*

$A : s, t \overset{d_1}{\Longrightarrow} u$ $\qquad$ $B : p \overset{d_2}{\Longrightarrow} \neg\mathrm{d}1$

$C : v \overset{d_3}{\Longrightarrow} p$ $\qquad$ $D : r, u \overset{d_4}{\Longrightarrow} \neg p$

The graphical representation of this argumentation system is displayed in Figure 10.

If we take $\mathbb{F} = \mathsf{Prem}$, then: $\mathsf{In}(u) = \{r, s, t\}$ and $\mathsf{Out}(u) = \{v\}$. If we take $\mathbb{F} = \mathsf{id}$, then: $\mathsf{In}(u) = \{A, D\}$. If we take $\mathbb{F} = \mathsf{SubConc}$, then: $\mathsf{In}(\neg\mathrm{d}1) = \{p\}$.

5.2.1 Sufficiency in ASPIC$^+$

As has been stated in Section 3.2.1, for abstract argumentation it was helpful to add a sufficiency requirement. Sufficiency aids with narrowing down the number of explanations by only looking at sufficient admissible extensions as possible explanations. This was also a manner to conceive a notion of relevance, to leave out irrelevant and unrelated arguments in the explanation. Sufficient explanations have been studied before in the setting of ASPIC$^+$ by Borg and Bex [BB21c]. They found that it can lead to a meaningful reduction in the size of an explanation, thus it is worthwhile to once again consider sufficiency for explanations in this research.

An interesting distinction to make here, is that in an abstract argumentation setting, one only considers the arguments in its entirety as one entity. In the structured setting of ASPIC$^+$, one does not only consider the arguments, but also the structure of those arguments. It is certainly possible that for the acceptance of argument A with conclusion ϕ, no specific argument is sufficient, but one must consider whether the arguments are build on the same sub-conclusion or premise. This would make that sub-conclusion or premise sufficient, rather than an entire argument. It is also possible that two arguments would be considered sufficient, but they are built on the same premise: that would make that single premise the sufficient explanation. Sufficiency can give different insights in the setting of ASPIC$^+$ compared to abstract argumentation.

With sufficiency already defined in Definition 3.6, it is possible to define new sufficient acceptance explanations in ASPIC$^+$:

Definition 5.10. Let $\mathcal{AF}(AT) = \langle Args, Att \rangle$ be an AF, based on AT and let $\phi \in \mathcal{L}$ be accepted. Then, sufficient acceptance explanations are defined by:

$$\mathsf{SuffIn}(\phi) = \{\mathbb{F}(\mathsf{SuffIn}(A)) \mid A \in Args \text{ and } \mathsf{Conc}(A) = \phi\}.$$

Sufficient non-acceptance explanations are defined by:

$$\mathsf{SuffOut}(\phi) = \{\mathbb{F}(\mathsf{SuffOut}(A)) \mid A \in Args \text{ and } \mathsf{Conc}(A) = \phi\}.$$

A sufficient acceptance explanation for a formula contains all the sufficient admissible extensions of which the formula is the conclusion of at least one of the arguments. A sufficient non-acceptance explanation for a formula contains all the sufficient admissible extensions of which all the arguments (indirectly) attack the arguments of which the formula is the conclusion and its defenders as possible explanations. The function $\mathbb{F}$ can be applied to both type of explanations.

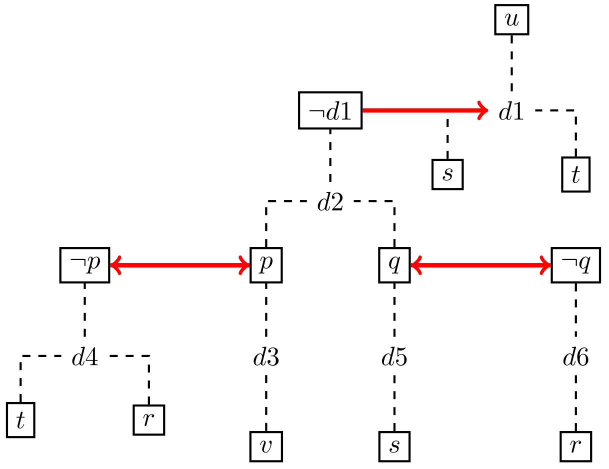

Figure 11: Another example argumentation system in ASPIC$^+$

Example 23. *Let $AS = \langle \mathcal{L}, \mathcal{R}, n \rangle$ where the rules in $\mathcal{R}$ are such that, with $\mathcal{K} = \mathcal{K}_n = \{r, s, t, v\}$ and with $\mathcal{R}_d = \{d_1, d_2, d_3, d_4, d_5, d_6\}$, the following arguments can be derived:*

$A : s, t \xRightarrow{d_1} u$ $\quad$ $B : p, q \xRightarrow{d_2} \neg d1$ $\quad$ $C : t, r \xRightarrow{d_4} \neg p$

$D : v \xRightarrow{d_3} p$ $\quad$ $E : r \xRightarrow{d_6} \neg q$ $\quad$ $F : s \xRightarrow{d_5} q$

The graphical representation of this argumentation system is displayed in Figure 11.

If we take $\mathbb{F}$ = id, then: SuffIn(u) $= \{\{A, C\}, \{A, E\}, \{A, C, E\}\}$. If we take $\mathbb{F}$ = Prem, then the possible sufficient acceptance explanation for u is: SuffIn(u) $= \{s, t, r\}$.

If we take $\mathbb{F}$ = Prem, then the possible sufficient non-acceptance explanations for $\neg d1$ are: SuffOut($\neg d1$) $= \{\{t,\ r\},\ \{r\}\}$. If we take $\mathbb{F}$ = id, then: SuffOut($\neg d1$) $= \{\{C\},\ \{E\},\ \{C,\ E\}\}$.

Example 23 shows the different explanations sufficiency can provide in ASPIC$^+$, based on how $\mathbb{F}$ is instantiated. Sufficiency in ASPIC$^+$ can thus provide new insights, compared to sufficiency in abstract argumentation.

5.3 Contrastiveness in ASPIC$^+$

Having defined the basic concepts of ASPIC$^+$, explanations and sufficient explanations in the previous sections, it is time to consider contrastive explanations. In this section, the types of contrasts will be handled one by one with examples to see how the adjusted definitions for ASPIC$^+$ will perform.

Just as before, the first contrast to discuss is the nonoccurence-contrast. Once again, the sufficient explanation for a nonoccurence-contrast will just provide the sufficient elements of arguments for the acceptance or non-acceptance of that formula.

Definition 5.11. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF, based on AT and let $\phi \in \mathcal{L}$ be accepted. Then, sufficient nonoccurence-contrast explanations are defined by:

$$\mathsf{ASPICNonContSuff}(\phi) = \{\mathbb{F}(\mathsf{SuffIn}(P)) \mid P \in Args \text{ and } \mathsf{Conc}(P) = \phi\}.$$

The definition provided here is the same as Definition 5.10. This is once again due to the nature of a nonoccurence contrast - to give a sufficient explanation for a formule ϕ, one needs to provide the elements of arguments that make it so when they are accepted, ϕ is accepted. This implies, by nature, that $\neg\phi$, the negation of ϕ, is not accepted. Thus it is satisfactory to just provide the sufficient arguments for the formula as a sufficient contrastive explanation for the nonoccurence-contrast. In the case of explanations within ASPIC$^+$, the function $\mathbb{F}$ can be instantiated in various ways to provide various possible explanations (that is, for example in terms of premises, arguments or sub-conclusions).

Example 24. *Consider again the argumentation system displayed in Figure 10. If we take $\mathbb{F} = \mathsf{Prem}$, then, when asking for a sufficient contrastive explanation for "Why u rather than $\neg u$?" one receives: $\mathsf{ASPICNonContSuff}(u) = \{s, t, r\}$ as the sufficient explanation.*

The second contrast to look into is the property-contrast. By taking the example contrastive question *"Why ϕ rather than ψ?* one explains why formula ϕ is accepted while formula ψ is not. To provide an explanation in ASPIC$^+$, the buildup is similar to the previous definition, and once again makes use of the function $\mathbb{F}$. This time however, there are two formulas to be considered:

Definition 5.12. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF, based on AT and let $\phi \in \mathcal{L}$ be accepted. Then, sufficient property-contrast explanations are defined by:

$$\mathsf{ASPICPropContSuff}(\phi, \psi) = \{\mathbb{F}(S \setminus \bigcup \mathsf{SuffIn}(B)) \mid S \in \mathsf{SuffIn}(A),\ A, B \in Args,\ \mathsf{Conc}(A) = \phi \text{ and } \mathsf{Conc}(B) = \psi\}.$$

To find the appropriate contrastive explanation, one takes the set difference over the acceptance explanations for all arguments A of which the conclusion is ϕ, with the union of all possible acceptance explanations for an argument B with conclusion ψ. This leaves one with the explanations for all arguments A with conclusion ϕ, excluding any possible elements of arguments that could cause an acceptance for an argument B with conclusion ψ.

Example 25. *Consider the argumentation system displayed in Figure 11. If we take $\mathbb{F} = \mathsf{Prem}$, then, when asking for a sufficient explanation for "Why u rather than q?", one starts with the sufficient explanation for u (which is $\{s, t, r\}$) and takes the set difference with the sufficient explanation for q ($\{s\}$). This results in: $\mathsf{ASPICPropContSuff}(u, q) = \{t, r\}$ as the possible sufficient explanation.*

If we take $\mathbb{F} = \mathsf{id}$ for the same question, then $\mathsf{ASPICPropContSuff}(u, q)$ results in the explanations $\{\{A, C\}, \{A, E\}, \{A, C, E\}\} \setminus \{F\} = \{\{A\}, \{A, E\}, \{A, C, E\}\}$.

The final contrast to discuss is the object-contrast. For this contrast, one wants to find the difference between situations where arguments A with conclusion ϕ are accepted and arguments B with conclusion ψ are not, and situations where both arguments with their respective conclusions are accepted. The definitions in ASPIC$^+$ for a sufficient explanation for the object-contrast is as follows:

Definition 5.13. Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF, based on AT and let $\phi \in \mathcal{L}$ be accepted. Then, sufficient object-contrast explanations are defined by:

$$\mathsf{ASPICObjContSuff}(\phi, \psi) = \{\mathbb{F}(S \setminus \bigcup(\mathsf{SuffIn}(A) \cap \mathsf{SuffIn}(B))) \mid S \in \mathsf{SuffIn}(A) \cap \mathsf{SuffOut}(B),\ A, B \in Args,\ \mathsf{Conc}(A) = \phi \text{ and } \mathsf{Conc}(B) = \psi\}.$$

This definition once again follows the previously established pattern: one takes the set difference between the admissible extensions in which arguments A with conclusion ϕ are accepted and arguments B with conclusion ψ are not, with the admissible extensions in which both arguments A and B are accepted (where ϕ is the conclusion of the arguments A and ψ the conclusion of the arguments B).

Example 26. *Consider again the argumentation system displayed in Figure 11. If we take $\mathbb{F} = \mathsf{Prem}$, then, one could ask for a sufficient explanation for the question "Why u and not $\neg p$ rather than u and $\neg p$?". To answer this question, first one finds the sufficient explanations for "Why u and not $\neg p$?", which results in the set $\{s, t, r, v\}$. Now one takes the set difference with the union of explanations for "Why u and $\neg p$?", which is $\{s, t, r\}$. This results in $\mathsf{ASPICObjContSuff}(u, \neg p) = \{v\}$.*

5.4 Evaluation of Contrastiveness in ASPIC$^+$

In the previous sections, definitions for various types of contrastiveness have been constructed and applied in a ASPIC$^+$ setting. So far, these definitions seem to perform well in the various examples that have been provided and were shown to be capable of giving adequate explanations. As was stated in 5.1, ASPIC$^+$ offers a rich set of expressive features, such as support for different types of attacks (undermining, undercutting and rebuttal) and preferences. These expressive capabilities enable a more nuanced representation of arguments and their interactions, allowing for more detailed and precise explanations. These features contribute to the clarity, precision, and comprehensibility of the explanations generated using ASPIC$^+$, which can not be achieved in abstract argumentation.

Just as was done for the definitions contrastiveness in preference-based argumentation in Section 4.4, the definitions of contrastiveness within ASPIC$^+$ need to be evaluated. Once again, the topic of contrastiveness based on findings in literature from the social sciences and humanities has not yet been studied within an ASPIC$^+$ setting, thus the manner of evaluation will follow the same process as in Section 4.4. As was stated in Section 5.1, ASPIC$^+$ can be used in various domains where argumentation and reasoning play a crucial role, such as law and policy-making [OBBT22]. This makes the previously used example situations well-suited for this evaluation as well. That is why, in this section, the same real-world problems will be cited (based on [BPS09], [Pra15] and [BB21b]) and the definitions of contrastiveness in ASPIC$^+$ will be employed to see whether they will come up with satisfactory explanations to the various situations.

The first example will be based on the case of *Keeble v Hickeringill* (1707). Now, it is important to first adjust the arguments so that they are suitable for the structured argumentation setting of ASPIC$^+$.

Example 27. *Let $AS = \langle \mathcal{L}, \mathcal{R}, n \rangle$ where the rules in $\mathcal{R}$ are such that, with $\mathcal{K} = \mathcal{K}_n = \{li, di, ca, hu, po\}$, the following arguments can be derived:*

$$A : po \overset{d_1}{\Longrightarrow} se \qquad B : di, \neg ri \overset{d_2}{\Longrightarrow} \neg d1 \qquad C : li \overset{d_3}{\Longrightarrow} \neg d2$$
$$D : ca \overset{d_4}{\Longrightarrow} \neg ri \qquad E : hu \overset{d_5}{\Longrightarrow} ri$$

The atoms represent the following premises and conclusions:

li: Plaintiff is pursuing his own livelihood.

di: Defendant disrupts plaintiffs business by scaring away the ducks.

ca: The ducks were not yet caught by anyone.

hu: Plaintiff was hunting on his own land.

po: Plaintiff owns a pond where he can attract ducks.

se: Plaintiff wants to sell caught ducks in owned pond for a living.

ri: Plaintiff has rights to the ducks.

$\neg ri$: Plaintiff has no right to the ducks.

$\neg d1$: Legal certainty is promoted.

$\neg d2$: Protection of property is promoted.

The graphical representation of this argumentation system is displayed in Figure 12.

Now that this situation has been sketched out in an ASPIC$^+$ setting, it is possible to evaluate the definitions for contrastiveness in ASPIC$^+$, and see how they perform and what kind of explanations would be given in this situation - to then also be able to compare what ASPIC$^+$ adds to an explanation, and why it is of importance to be studied.

Example 28. *Consider the case Keeble v Hickeringill pictured in Figure 12. One could ask the question "Why does the plaintiff have right to the ducks, rather than not having rights to them?". In terms of the argumentation system, this question would "Why ri, rather than $\neg ri$?". In this case, the type of contrast is a nonoccurence-contrast.*

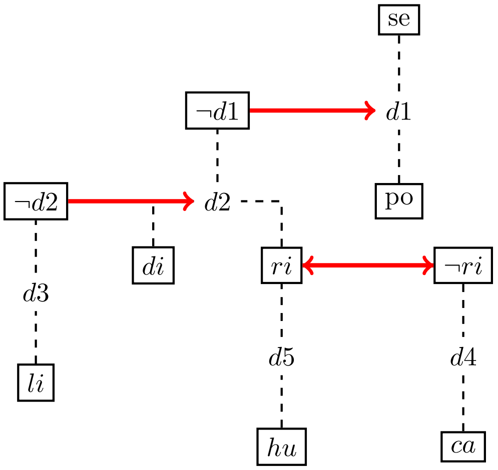

Figure 12: argumentation system in ASPIC$^+$ of the case Keeble v Hickeringill

If we take $\mathbb{F} = \mathsf{Prem}$*, then, when asking "Why* ri *rather than* $\neg ri$*?", one receives the explanation:* $\mathsf{ASPICNonContSuff}(ri) = \{hu\}$ *as possible sufficient explanation. In natural language, the explanation would thus be "Because the plaintiff was hunting on his own land."*

Example 29. *Consider again the case Keeble v Hickeringill pictured in Figure 12. One could ask for an explanation in response to the result of the court ruling: "Why is the plaintiff allowed to sell ducks caught in the pond owned by him for a living, rather than promoting legal certainty?". In terms of the argumentation system, this question would be: "Why se, but not* $\neg$*d1?". In this case, the type of contrast is a property-contrast.*

If we take $\mathbb{F} = \mathsf{Id}$*, then, when asking "Why se, but not* $\neg$*d1?", one starts with the sufficient explanations for se (which are* $\{\{A, C\}, \{A, D\}, \{A, D, E\}\}$ *and takes the set difference with the sufficient explanation for* $\neg$d1 $(\{\emptyset\})$*. This results in:* $\mathsf{ASPICPropContSuff}(p, \neg\mathrm{d}1) = \{\{A, C\}, \{A, D\}, \{A, D, E\}\}$ *as the three possible sufficient explanations.*

Example 30. *Consider again the case Keeble v Hickeringill pictured in Figure 12. One could ask for an explanation for the following question: "Why is the plaintiff allowed to sell ducks caught in the pond he owns while the plaintiff has no right to the ducks, rather than allowing to sell ducks caught in the pond he owns while having right to the ducks?". In terms of the argumentation system, this question would be: "Why se and not* ri*, rather than se and* ri*?". In this case, the type of contrast is an object-contrast.*

If we take $\mathbb{F} = \mathsf{Prem}$*, then, one could ask the question "Why se and not* ri*, rather than se and* ri*?". To answer this question, first one finds the sufficient explanations for "Why se and not* ri*", which results in* $\{li, ca, po\}$*. Now one takes the set difference with the union of explanations for "Why se and* ri*?", which is* $\{li, hu, po\}$*. This results* $\mathsf{ASPICObjContSuff}(se, ri) \in \{ca\}$*.*

Examples 28, 29 and 30 show that my definition for the nonoccurence-, property- and object-contrast in ASPIC$^+$ perform well in that they are able to provide the relevant premises for a sufficient explanation. All three contrasts are evaluated in this first example case. The resulting explanations are interesting to compare to the previous settings of abstract argumentation, as one can specify wanting, for example, to have premises as the explanation. More specific situations can be modelled in ASPIC$^+$, which allows for more context and more information while modelling the same case. In order to properly reflect this, it is valuable to be able to provide explanations in ASPIC$^+$. This leads to a type of explanation that could not be acquired in an abstract argumentation setting (which was also already illustrated in Example 25).

To further elaborate on this, the next evaluation will regard the case of Olga Monge, as it was described in Section 4.4. This case has been studied in an ASPIC$^+$ setting before, so for this evaluation, the same set-up as in [Pra15] will be used.

Example 31. *Let $AS = \langle \mathcal{L}, \mathcal{R}, n \rangle$ where the rules in $\mathcal{R}$ are such that, with $\mathcal{K} = \mathcal{K}_n = \{ru, rb, bg, wi, nr, pg, pr, ma\}$, the following arguments can be derived:*

$$A : wi, nr \overset{d_1}{\Longrightarrow} fi \qquad B : ru, ol \overset{d_2}{\Longrightarrow} nr \qquad C : rb, bg \overset{d_3}{\Longrightarrow} ol \qquad D : nr, pg \overset{d_4}{\Longrightarrow} es$$
$$E : es, pr \overset{d_5}{\Longrightarrow} \neg ol. \qquad F : \neg ol, ru \overset{d_6}{\Longrightarrow} fu. \qquad F : fu, ma \overset{d_7}{\Longrightarrow} \neg fi$$

The atoms represent the following premises and conclusions:

ru: If we should adopt Rule R as the valid rule, then Rule R.

rb: The Old Rule makes that employers can run their businesses as they see fit.

bg: Employers being able to run their businesses as they see fit is good.

wi: Monge was employed at will.

nr: The new rule makes that employees cannot be fired in bad faith, malice or retaliation.

pg: If employees cannot be fired in bad faith, malice or retaliation then the economic system and the public good is promoted.

pr: Promoting the economic system and the public good is good.

ma: Monge was fired in malice.

fi: Monge could be fired for no reason.

ar: Someone employed at will can be fired for any reason or no reason at all.

ol: We should adopt the Old Rule as the valid rule.

np: The new rule promotes the economic system and the public good.

fu: Someone employed at will can be fired for any reason or no reason at all unless the employee is fired in bad faith, malice or retaliation.

$\neg ol$: We should adopt the New Rule as the valid rule.

$\neg fi$: Olga Monge could not be fired for no reason.

The graphical representation of this argumentation system is displayed in Figure 13 as pictured in [Pra15].

Now that this second case has been sketched out in an ASPIC$^+$ setting, it is possible to continue with the evaluation of the definitions for contrastiveness in ASPIC$^+$. As one can gather, even though it is the same case, the argumentation system pictured in Figure 13 is more elaborate than the same case pictured in Figure 6. Certain information can only be modelled in ASPIC$^+$. For example, premise ru (*"If we should adopt Rule R as the valid rule, then Rule R"*) is applied in both argument B and argument F. This dual usage of the same premise for two different arguments is something that could not be modelled in abstract argumentation.

Example 32. *Consider the case Olga Monge v Beebe Rubber Company pictured in Figure 13. One could ask the question "Why could Olga Monge not be fired for no reason, rather than it being possible?". In terms of the argumentation system, this question would "Why $\neg fi$, rather than not $\neg fi$?" In this case, the type of contrast is a nonoccurence-contrast.*

If we take $\mathbb{F}$ = Prem, then, one could ask for a sufficient explanation for the question "Why $\neg fi$ rather than not $\neg fi$?". This results in: ASPICNonContSuff($\neg fi$) = $\{nr, pg, pri, ru, ma\}$ as the sufficient explanation.

If we take $\mathbb{F}$ = SubConc, then, one could ask for a sufficient explanation for the question "Why $\neg fi$ rather than not $\neg m$?". This results in: ASPICNonContSuff($\neg fi$) = $\{es, \neg ol, fu\}$ as the sufficient explanation. In natural language, this would come down to: "Because the new rule promotes the economic system and the public good, we should adopt the new rule as the valid rule, which disallows Olga Monge to be fired for no reason".

In Example 32 once again the definition nonoccurence-contrast has performed well, as it was able to find the sufficient premises and sub-conclusions as explanations for the contrast. There are a few difference with this same situation in PAFs. The first, and most obvious, is that while this is

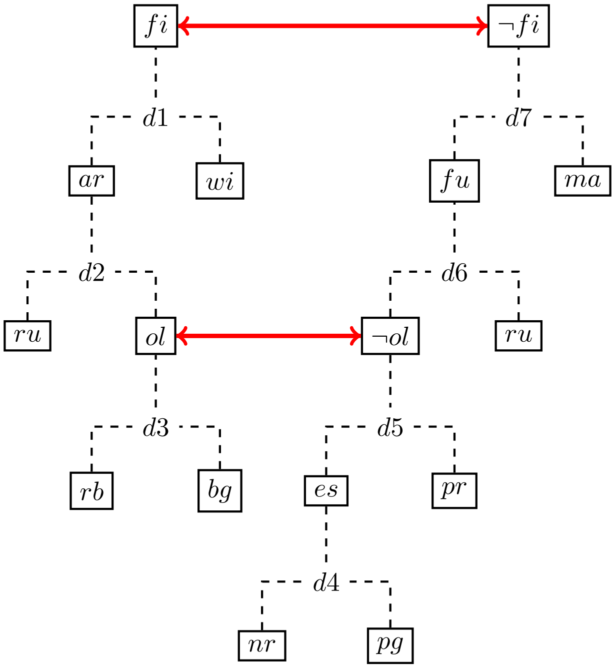

Figure 13: argumentation system in ASPIC$^+$ of the case Olga Monge v Beebe Rubber Company from [Pra15]

the same case, one can see how much much more information, and with that, depth, can be added to the argumentation structure in ASPIC$^+$ compared to the argumentation framework in abstract argumentation. The second is that in ASPIC$^+$, one is able to specify what kind of explanation is wanted - whether one wants the conclusions, premises, or sub-conclusions. These two differences make way for more detailed explanations than an abstract argumentation framework could provide. This will be further illustrated at the hand of the final evaluation example.

As a final evaluation, the example situation provided by [BB21b] will be considered again. This evaluation was chosen as it is a more recent case about possible malafide webshops, based on application at the Netherlands Police.

Example 33. *Let $AS = \langle \mathcal{L}, \mathcal{R}, n \rangle$ where the rules in $\mathcal{R}$ are such that, with $\mathcal{K} = \mathcal{K}_n = \{cf, rc, sa, ka, kp, rr\}$. The used atoms mean the following: From these atoms and their negations, the following arguments can be derived:*

$$
\begin{array}{llllll}
A_1 : cf & A_2 : rc & A_3 : sa & A_4 : ka & A_5 : kp & A_6 : rr \\
B_1 : A_1 \overset{d_1}{\Longrightarrow} iw & & B_2 : A_2 \overset{d_2}{\Longrightarrow} \neg n(d_1) & & B_3 : A_5 \overset{d_5}{\Longrightarrow} \neg rc & \\
B_4 : B_1, A_3 \overset{d_3}{\Longrightarrow} m & & B_5 : A_4 \overset{d_4}{\Longrightarrow} \neg n(d_3) & & B_6 : A_6 \overset{d_6}{\Longrightarrow} \neg ka &
\end{array}
$$

The provided atoms represent the following premises and conclusions:

cf: a complaint was filed.

m: the webshop is malafide.

iw: an investigation is done.

sa: the url is suspicious.

rc: the complaint is retracted.

kp: the webshop owner is known by the police.

ka: the address is registered at the chamber of commerce.

rr: the registration was recently retracted.

See Figure 14 for a graphical representation of the corresponding framework.

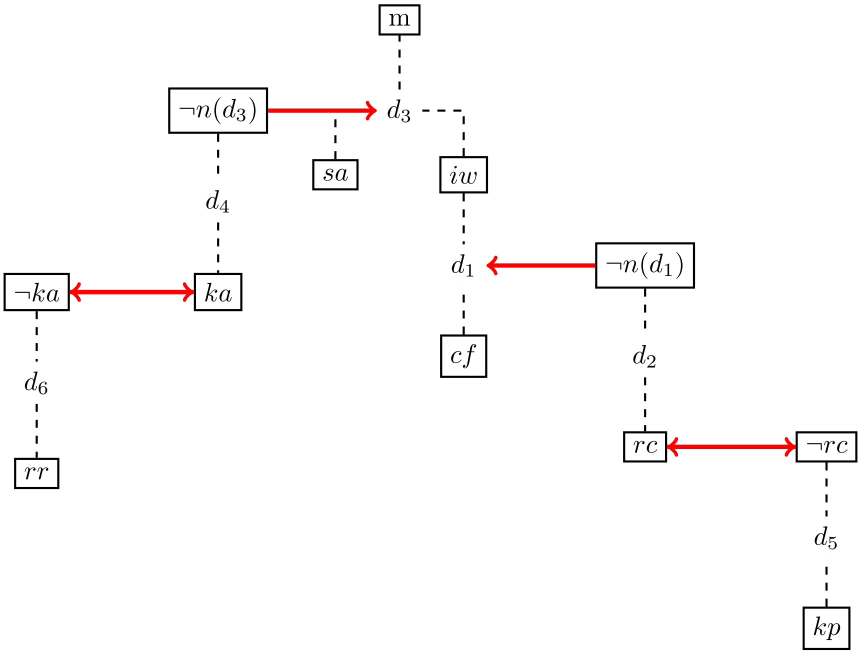

Figure 14: argumentation system in ASPIC$^+$ of the possible malafide workshops from [BB21b]

Example 34. *Consider the scenario displayed in Figure 14 based on malafide webshop cases at the Netherlands Police. One can ask the contrastive question "Why is an investigation done when the address is not registered, rather than doing an investigation when the address is registered?". In terms of the AS displayed, this question concerns atoms iw and ka: "Why iw and not ka, rather than iw and ka?". The type of contrast to consider here is an object-contrast.*

If we take $\mathbb{F} = \mathsf{Prem}$, then, one can ask the question "Why iw and not ka, rather than iw and ka?". To answer the question, first one finds the sufficient explanations for "Why iw and not ka?", which results in $\{cf, kp, rr\}$. Now one takes the set difference with the union of explanations for "Why iw and ka?", which is $\{cf, kp, ka\}$. This results $\mathsf{ASPICObjContSuff}(iw, ka) = \{rr\}$. In other words, the explanation to the question would be because the registration was recently retracted.

At the hands of Examples 28 to 34, it has been shown that the situations modelled in Section 4.4 in abstract argumentation can be modelled with more detail and additional information in ASPIC$^+$. To correctly reflect this new layer of depth to the structures, my contrastive explanations in ASPIC$^+$ can provide the additional information that explanations in other argumentation approaches can not provide. Depending on how the function $\mathbb{F}$ is instantiated, explanations could be given at the hand of premises, conclusions or sub-conclusions. This is not possible in abstract argumentation, where only the abstract argument is considered. Being able to do this in ASPIC$^+$ adds a new layer and function to explanations, that is impossible to achieve in abstract argumentation approaches, and much more comparable to real-life argumentation. As has been mentioned in Section 5.1, the use of ASPIC$^+$ comes with many advantages, such as expressiveness and flexibility. It is capable of providing new information, and more precise information as well. This is why in real life, applications based on ASPIC$^+$ are employed (such as at the Netherlands Police, which employs several applications based on structured argumentation frameworks [OBBT22]). For this reason, it is vital to study explanations in ASPIC$^+$, to make such applications explainable to people. Especially contrastive explanations are important to study in this setting, as they have been shown to be better tailored to more specific explanations.

This concludes the section on ASPIC$^+$. In this section, it has been shown that all of the three contrasts (nonoccurence-, object- and property-contrast) can be adjusted and still work even in a structured argumentation setting such as ASPIC$^+$. By creating various adjustments and new definitions to account for ASPIC$^+$'s structured nature, my contrastive definitions still perform well and can be used to find suitable contrastive explanations, while adding valuable insights that can only be achieved by studying contrastive explanations in ASPIC$^+$.

6 Discussion

In this section, first the subquestions of the main research question will be answered at the hand of the work done in the previous sections. After this has been done, my findings for contrastive explanations will be evaluated at the hand of the value measures introduced in Section 2.4. Then, the implications of this work will be discussed, to show why these results matter. Finally, this section will end with discussing some of the limitations of this work, which can be avenues for further studies.

6.1 Answering the Research Subquestions

In this thesis, the topic of contrastiveness within the field of argumentation has been studied thoroughly. Throughout all the previous sections, the various subquestions of the main research question have been answered, which will be discussed in this subsection. The main research question is:

> *Based on different definitions given in existing social sciences literature, how can contrastiveness be modelled in formal argumentation?*

To answer this question, first, in Section 2.1 existing literature from the fields of humanities and the social sciences about the topic of contrastiveness was summarised and compared against one another. This was done extensively to answer the first subquestion:

1. *What are various definitions given by the fields of humanities and the social sciences for contrastiveness and what similarities can be concluded?*

From this literature research, it was found that while many researchers had varying opinions on what contrastiveness entailed, some similarities and agreements could be found between the existing research. Summarising the findings, the definition of contrastiveness was split up into four different types of contrast: the nonoccurence-contrast, the object-contrast, the property-contrast and the over time-contrast. Aside from the contrasts, the two properties of relevance and compatibility between fact and foil were established. Fact and foil must always be relevant to one another, and fact and foil are incompatible for the nonoccurence-, property- and over time-contrast. For the object-contrast, fact and foil may be compatible. Finally, the requirements of contrastive questions were established: that of truth value, type of contrast and that there must only be one contrast in a contrastive question.

From these findings, it was possible to start modelling the four types of contrasts in abstract argumentation, to answer the second subquestion:

2. *How can these definitions of contrastiveness be modelled in an abstract argumentation setting?*

This was done in Section 3. While researching the answer for this question, it was found that the over time-process would be difficult to model in a static abstract argumentation setting. If one wanted to properly model a setting where time has passed, a non-static argumentation framework (such as IAFs) would be required. This was deemed out of the scope of this research, so only the other three contrasts (nonoccurence, object and property) were modelled in an abstract argumentation setting. The explanations were modelled by using admissible extensions, and sufficiency was used to narrow down the big explanations admissible extensions could provide.

Once the second subquestion was answered, it was important to see how the different contrasts could be modelled in other settings. The abstract argumentation frameworks of Dung can be seen as a simple base, which can be expanded upon. Various of such argumentation approaches exist. To show that my definitions of contrastiveness can be applied in more than just the setting of Dung's abstract argumentation frameworks, the following question was to be answered:

3. *How can these definitions be applied and evaluated in preference-based argumentation?*

In Section 4, first, the value of adding preferences to argumentation was expanded upon. Then, my definitions for contrastiveness were modelled in a preference-based argumentation setting. Finally, those definitions were evaluated at the hand of real-life example situations to show that they perform well and provide explanations which can account for the influence of preferences.

Abstract argumentation, however, does come with some limitations. In the real world it has been shown that structured argumentation is more preferred for XAI applications, and various of such applications that are currently in use have been based on ASPIC$^+$. Thus, the following subquestion came to be:

4. *How can these definitions be applied and evaluated in ASPIC$^+$?*

This final subquestion was answered in Section 5, and the same set-up was used as the previous section on PAFs. First, the relevance of ASPIC$^+$ was elaborated upon. Then, the definitions of contrastiveness were adjusted to account for the structured setting of ASPIC$^+$. Once this was done, the definitions were evaluated at the hand of various real world examples, and once again it was shown that my definitions provide explanations which account for the structured argumentation setting of ASPIC$^+$.

By summarising the findings of this thesis, it has been shown that the subquestions have been answered: this thesis has based its definitions of contrastiveness on existing literature in the field of humanities and the social sciences, and shown how contrastiveness can be modelled in formal argumentation - both abstract and structured argumentation, and then continued to evaluate each of these findings.

6.2 Evaluation of Contrastive Explanations

In Section 2.4, five values were named to measure how 'good' or 'powerful' a contrastive explanation is. These five measures are based on the findings by Ylikoski and Kuorikoski [YK10]. In this subsection, the findings of this thesis will be evaluated at the hand of these measures.

The first value is that of *non-sensitivity*: explanations should be independent of changes within the background conditions. All of my contrastive explanations can be said to be non-sensitive, as they all account for a notion of relevance. In abstract argumentation (in both Dung's frameworks and PAFs) and ASPIC$^+$, relevance is modelled through sufficiency. Sufficiency is defined in Definition 3.6 in such a manner that it only considers relevant arguments. By having this notion of relevance, explanations will not be changed by having some unrelated background argument change - only the arguments that are relevant to the explanation will influence the explanation. Thus, my contrastive explanations are non-sensitive, as they are independent of changes within unrelated arguments.

The second value is that of *precision*: how precisely the explanation characterises the explanandum. The more detailed an explanation is, the better it generally is. In this case as well, I would argue that my contrastive explanations are precise. By making use of sufficiency, in all settings the explanations can pinpoint which arguments truly matter for the explanandum. However, it can be said that in the setting of ASPIC$^+$, the explanations are more precise than in an abstract argumentation setting. This makes sense, as in such a structured setting one can specify which part of the arguments one would like as the explanation. This would make contrastive explanations in ASPIC$^+$ more precise than in an abstract argumentation setting, and thus score higher on that value.

The third value Ylikoski and Kuorikoski mentioned is that of *factual accuracy*: is the explanation true? This one is more difficult to objectively measure, as in this thesis abstract and structured argumentation are used, where the content of the arguments is not always specified. Abstract argumentation states nothing about a truth value of the arguments. Preferences however, could create a higher factual accuracy in abstract argumentation, by providing preference relations in which the factual arguments are preferred. PAFs can thus be said to provide a higher factual accuracy. In ASPIC$^+$, one can make a clear distinction between defeasible and strict arguments and rules. Defeasibility implies that something does not have to be true - it can be defeated by arguments, while strict arguments are based on logical inference and thus must hold. So in this case, one could say that contrastive explanations in ASPIC$^+$ can account for a higher factual accuracy than in abstract argumentation, but preferences can increase the factual accuracy of contrastive explanations in abstract argumentation.

The fourth value is that of the *degree of integration*: is it consistent with a well-supported theory? This one is harder to evaluate, as once again by working with formal argumentation the content of arguments is not given. While my definitions for contrastiveness and explanations *are* based on previous research and established theories, the content of the contrastive explanations themselves is entirely context-dependent, which is something that is not always provided in formal argumentation.

The fifth and final value is that of *cognitive salience*: the ease with which the reasoning on an explanation can be followed. This is possibly one of the most important values, as an explanation should be possible to follow for anyone to make it usable. At the hand of examples in Sections 3.3, 4.4 and 5.4, I have shown that these explanations are easy to follow. The examples illustrate how the modelled contrastiveness can find contrastive explanations in the three frameworks of abstract argumentation, preference-based argumentation and ASPIC$^+$, thus I argue that my explanations have a high cognitive salience.

6.3 Implications

The first findings in this research, the definitions of contrastiveness found in existing literature, build on existing research by Borg and Bex [BB22a]. The types of contrasts that were concluded from the literature research are comparable to their findings, and so is the way they are modelled in abstract argumentation. There do exist various differences between the definitions (elaborated upon in Section 3.4), however, it can still be said that the results of the first and second subsection can be seen to validate their work, and imply that this is a good and substantiated way to model contrastiveness. By doing my own research, similar results were found. This implies that these various notions of contrast are good grounds for further research into contrastiveness.

While previous research has focused on contrastiveness in abstract argumentation, these results demonstrate that it is also worthwhile to study contrastiveness in other argumentation approaches, such as preference-based argumentation frameworks and ASPIC$^+$. To the best of my knowledge, contrastiveness based on findings in the social sciences and humanities, which resulted in various types of contrast, had not yet been studied within these other approaches, and this research is the first to do so. Preference-based argumentation frameworks and ASPIC$^+$ both come with many advantages, and the first and foremost is that they both are capable of better emulating human-like argumentation. In the case of PAFs, being able to account for preferences makes the abstract arguments more understandable to humans, as they are capable of providing more context in an abstract setting, and they allow for more subjective judgements to be made. As a result of this, PAFs allow for much more flexibility and can provide different interpretations to the same situation. Preferences are inherent to the real world, and thus being able to account for them can only result more understandable explanations. Combining contrastive explanations, which have been shown to be the preferred way of explaining ([Mil19]), with preferences can only lead to a better application of XAI.

The same can be said for ASPIC$^+$, which has been shown to already have applications based on it that are currently in use for decision-making ([OBBT22]). ASPIC$^+$ provides a structured approach which includes higher levels of expressiveness and flexibility, and thus is very well-suited for analyzing and evaluating arguments. Once again, combining this approach with contrastiveness allows for finding explanations that are highly suited for XAI.

The results of this research have shown that contrastiveness can give good explanations in the setting of abstract argumentation, preference-based abstract argumentation and ASPIC$^+$, and should be taken into account when considering how to improve contrastive explanations in formal argumentation for use in (argumentative) XAI.

6.4 Limitations

Due to the scope of this project, some of the findings are not final and there are some limitations present. This section discusses the limitations of this research.

The first limitation to mention is that this work does not account for the over time-contrast. Even though in multiple literary sources this contrast could be found, due to time constraints, this work does not further elaborate on it. Frameworks such as IAFs would be an interesting approach to research more into the over time-contrast. As multiple past research has stated this to be a unique contrast ([DB08], [BW02], [Mil21]) which differs from the other contrasts, it would be worthwhile for future research to find a good way to define this contrast as well.

The second limitation is a rather specific definition-bound limitation, concerning the object-contrast in preference-based argumentation. The way it is defined could be considered to be slightly odd; if there is a preference relation as a result of which the attack on the foil is successful, there is no admissible extension in which fact and foil are both true. This results in a noneffective taking of a set difference, as it would be taken with the empty set. This is the case in Example 16. While I do argue that my definition works and is capable of giving explanations that provide the relevant preferences, as it provides the preference relation that causes this non-acceptance of the foil in all admissible extensions, I can also say that there might be some space for improvements as well, or at least to further evaluate it in more different situations to see how this definition behaves.

The third limitation is regarding necessity within ASPIC$^+$. While necessity can be considered too strict for an abstract argumentation setting, it is important to note that within ASPIC$^+$, one does not only look at the arguments, but also at the structure of the arguments. It is possible that for the acceptance of one argument no argument is necessary, but one must consider whether those arguments are built on the same rule or premise. This would make that rule or premise necessary. However, the choice was made in this work to only continue with sufficient explanations. This decision came to be as a result of the manner in which my explanations are defined: the explained argument is always included in the explanation. For necessary arguments, this can get confusing - if one wants an acceptance explanation for argument A, and argument A has no necessary arguments, only sufficient arguments, rather than giving an empty set as a necessary explanation, the set $\{A\}$ would be returned. This is undesirable. Only by changing the basis of my explanation definitions would I be able to account for such situations. Due to the scope of this project, however, I did not have the time to research such alternatives. Nonetheless, it would be remiss not to consider necessity within ASPIC$^+$, as necessity and sufficiency can result in different explanations, and both can give different insights in the setting of ASPIC$^+$. That is why it would be worthwhile for future research to look into the topic and find a manner to model necessary contrastive explanations in ASPIC$^+$ as well.

The final limitation is the manner of evaluation. Contrastiveness has not yet been studied much in PAFs and ASPIC$^+$ before this work, and thus finding a way to properly evaluate it is vital. In this research, the various definitions of contrastiveness have been evaluated at the hand of providing examples based on some real-world situations. These evaluations showed that my definitions perform well and can provide good explanations in such situations. After this, the general notions of contrastive explanations have been evaluated at the hand of the values provided in [YK10]. To make these findings on contrastiveness more applicable to the real world, it might be better to try and create a proper application that make use of these definitions to more properly evaluate them, such as, for example, in PyArg [BO22]. This way, various situations can be easily adjusted and it can be seen how properly the definitions hold up in a much wider scale. It would also be easier to compare how the definitions change in the various settings (abstract and structured). So in short, while there has been a manner of evaluation present in this research which has shown that my definitions of contrastiveness provide effective explanations, one could only gain more by pursuing a more elaborate manner of evaluating.

7 Conclusion

The current research aimed to model contrastiveness in various formal argumentation settings. The main research question for this thesis was the following:

> *Based on different definitions given in existing social sciences literature, how can contrastiveness be modelled in formal argumentation?*

By doing an extensive literature research, it was concluded that contrastiveness can be divided into various notions of contrast; the nonoccurence-, property-, object- and over time-contrast. Out of these, three contrasts (nonoccurence, property and object) were modelled and defined in such a manner that they are applicable in abstract argumentation. By making some slight adjustments to the definitions, these same three contrasts were also modelled in preference-based argumentation, where the explanation also takes into account the relevant preferences, and the three contrasts were modelled in the structured argumentation setting of ASPIC$^+$, where they can account for different forms of explanations (premises, conclusions, etc.). By doing this, this thesis has shown how to model contrastiveness based on findings in the social sciences and the humanities in various settings of formal argumentation, and has applied these definitions at the hand of evaluations based on real world situations.

As has been stated before, to the best of my knowledge, this is the first work in which the various notions of contrastiveness found in literature have been formalised in multiple settings (abstract argumentation, PAFs and ASPIC$^+$), for use in argumentative XAI. This work is therefore a stepping stone for a broader discussion on contrastiveness in XAI and how to apply it in an effective manner. There still are some limitations to this work (elaborated upon in the previous section), and thus it might be worthwhile to further evaluate the definitions found here, study the over time-contrast, and see how they perform in various applications. Upon testing these definitions further and doing additional evaluations, some worthwhile notions might be found that can help either validate these definitions or improve them in some manner. Nonetheless, this work emphasises the importance and use of XAI in the present-day society, and provides a strong basis for the research of contrastiveness in not just abstract argumentation, but also in other approaches, which paves the way for the further study and application of contrastiveness in (argumentative) XAI.

References

[AB18] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[ABG+17] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.

[AC97] Leila Amgoud and Claudette Cayrol. Integrating preference orderings into argument-based reasoning. In *Proceedings of the International Conference on Qualitative and Quantitative Practical Reasoning (ECSQARU-FAPR)*, pages 159–170. Springer, 1997.

[AC98] Leila Amgoud and Claudette Cayrol. On the acceptability of arguments in preference-based argumentation. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1–7. UAI, 1998.

[AC02] Leila Amgoud and Claudette Cayrol. Inferring from inconsistency in preference-based argumentation frameworks. *Journal of Automated Reasoning*, 29:125–169, 2002.

[AL92] Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2):181–194, 1992.

[Ash17] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age.* Cambridge University Press, 2017.

[Bar94] Eric Barnes. Why P rather than Q? The curiosities of fact and foil. *Philosophical Studies*, 73(1):35–53, 1994.

[BB] AnneMarie Borg and Floris Bex. Improving explanations by integrating preferences.

[BB21a] AnneMarie Borg and Floris Bex. A basic framework for explanations in argumentation. *IEEE Intelligent Systems*, 36(2):25–35, 2021.

[BB21b] AnneMarie Borg and Floris Bex. Explaining arguments at the Dutch National Police. In Víctor Rodríguez-Doncel, Monica Palmirani, Michał Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor, editors, *AI Approaches to the Complexity of Legal Systems XI-XII (AICOL)*, pages 183–197. Springer, 2021.

[BB21c] AnneMarie Borg and Floris Bex. Necessary and sufficient explanations for argumentation-based conclusions. In *Proceedings of the 16th European Conference on Symbolic and Quantative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 45–58. Springer, 2021.

[BB22a] AnneMarie Borg and Floris Bex. Contrastive explanations for argumentation-based conclusions. In Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor, editors, *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1551–1553. IFAAMAS, 2022.

[BB22b] AnneMarie Borg and Floris Bex. Modeling contrastiveness in argumentation. In Floriana Grasso, Nancy L. Green, Jodi Schneider, and Simon Wells, editors, *Proceedings of the 22nd Workshop on Computational Models of Natural Argument (CMNA@COMMA 2022)*, volume 3205 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org, 2022.

[BBM+15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 2015.

[BCPV11] Trevor Bench-Capon, Henry Prakken, and Wietske Visser. Argument schemes for two-phase democratic deliberation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL)*, pages 21–30, 2011.

[BFG11] Robin D. Burke, Alexander Felfernig, and Mehmet H. Göker. Recommender systems: An overview. *AI Magazine*, 32(3):13–18, 2011.

[BGH$^+$14] Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014.

[BJN$^+$21] Dorothea Baumeister, Matti Järvisalo, Daniel Neugebauer, Andreas Niskanen, and Jörg Rothe. Acceptance in incomplete argumentation frameworks. *Artificial Intelligence*, 295:103470, 2021.

[BNRS18] Dorothea Baumeister, Daniel Neugebauer, Jörg Rothe, and Hilmar Schadrack. Verification in incomplete argumentation frameworks. *Artificial Intelligence*, 264:1–26, 2018.

[BO22] AnneMarie Borg and Daphne Odekerken. PyArg for solving and explaining argumentation in Python: Demonstration. In Francesca Toni, Sylwia Polberg, Richard Booth, Martin Caminada, and Hiroyuki Kido, editors, *Proceedings of 9th Conference on Computational Models of Argument (COMMA)*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, pages 349–350. IOS Press, 2022.

[BP10] Trevor J. M. Bench-Capon and Henry Prakken. A lightweight formal model of two-phase democratic deliberation. In Radboud Winkels, editor, *Proceedings of the 23rd Annual Conference on Legal Knowledge and Information Systems (JURIX)*, volume 223 of *Frontiers in Artificial Intelligence and Applications*, pages 27–36. IOS Press, 2010.

[BPS09] Trevor J. M. Bench-Capon, Henry Prakken, and Giovanni Sartor. Argumentation in legal reasoning. In Guillermo Ricardo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 363–382. Springer, 2009.

[Bre94] Gerhard Brewka. Reasoning about priorities in default logic. In Barbara Hayes-Roth and Richard E. Korf, editors, *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 940–945. AAAI Press / The MIT Press, 1994.

[BRT19] Pietro Baroni, Antonio Rago, and Francesca Toni. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning*, 105:252–286, 2019.

[BW02] Jeroen Van Bouwel and Erik Weber. Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, 32(4):437–449, 2002.

[Cay95] Claudette Cayrol. From non-monotonic syntax-based entailment to preference-based argumentation. In Christine Froidevaux and Jürg Kohlas, editors, *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU)*, volume 946 of *Lecture Notes in Computer Science*, pages 99–106. Springer, 1995.

[CDG$^+$15] Günther Charwat, Wolfgang Dvořák, Sarah A Gaggl, Johannes P Wallner, and Stefan Woltran. Methods for solving reasoning problems in abstract argumentation–a survey. *Artificial intelligence*, 220:28–63, 2015.

[CDLS07] Claudette Cayrol, Caroline Devred, and Marie-Christine Lagasquie-Schiex. Handling ignorance in argumentation: Semantics of partial argumentation frameworks. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 259–270. Springer, 2007.

[CH19] Miruna-Adriana Clinciu and Helen Hastie. A survey of explainable AI terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI)*, pages 8–13. Association for Computational Linguistics, 2019.

[Chu17] Michael Chui. Artificial intelligence: the next digital frontier? *McKinsey and Company Global Institute*, 47(3.6), 2017.

[CLS05] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Proceedings of the 8th European Conference of Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 378–389. Springer, 2005.

[CML00] Carlos Iván Chesñevar, Ana Gabriela Maguitman, and Ronald Prescott Loui. Logical models of argument. *ACM Computation Survey*, 32(4):337–383, 2000.

[CPB17] Seth Chin-Parker and Alexandra Bradner. A contrastive account of explanation generation. *Psychonomic Bulletin & Review*, 24(5):1387–1397, 2017.

[CRA+21] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4392–4399. International Joint Conferences on Artificial Intelligence Organization, 2021.

[DB08] Mark Day and George S Botterill. Contrast, inference and scientific realism. *Synthese*, 160(2):249–267, 2008.

[DBH18] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[Dun95] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[FT15] Xiuyi Fan and Francesca Toni. On explanations for non-acceptable arguments. In Elizabeth Black, Sanjay Modgil, and Nir Oren, editors, *Proceedings of the 3rd International Workshop on Theory and Applications of Formal Argumentation (TAFA)*, volume 9524 of *Lecture Notes in Computer Science*, pages 112–127. Springer, 2015.

[Gar80] Peter Gardenfors. A pragmatic approach to explanations. *Philosophy of Science*, 47(3):404–423, 1980.

[Gar82] Alan Garfinkel. Forms of explanation: Rethinking the questions in social theory. *British Journal for the Philosophy of Science*, 33(4), 1982.

[Gri75] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[Har65] Gilbert H. Harman. The inference to the best explanation. *Philosophical Review*, 74(1):88–95, 1965.

[Hes88] Germund Hesslow. The problem of causal selection. In Denis J. Hilton, editor, *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, pages 11–32. New York University Press, 1988.

[Hil90] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990.

[Hit99] Christopher Hitchcock. Contrastive explanation and the demons of determinism. *British Journal for the Philosophy of Science*, 50(4):585–612, 1999.

[HM19] Jörg Hoffmann and Daniele Magazzeni. Explainable AI planning (XAIP): overview and the case of contrastive explanation (extended abstract). In Markus Krötzsch and Daria Stepanova, editors, *Reasoning Web. Explainable Artificial Intelligence*, volume 11810 of *Lecture Notes in Computer Science*, pages 277–282. Springer, 2019.

[Lew86] David Lewis. Causal explanation. *Philosophical Papers*, 2:214–240, 1986.

[Lip90] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

[Lom10] Tania Lombrozo. Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, 61(4):303–332, 2010.

[LS89] Fangzhen Lin and Yoav Shoham. Argument systems: A uniform basis for nonmonotonic reasoning. *KR*, 89:245–255, 1989.

[MIBD02] Bernard Moulin, Hengameh Irandoust, Micheline Bélanger, and Gaëlle Desbordes. Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review*, 17(3):169–222, 2002.

[Mil19] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Mil21] Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[Mit97] Tom M. Mitchell. *Machine learning, International Edition.* McGraw-Hill Series in Computer Science. McGraw-Hill, 1997.

[MK93] Ann McGill and Jill Klein. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64(6):897, 1993.

[Moo06] James Moor. The Dartmouth college artificial intelligence conference: The next fifty years. *AI Magazine*, 27(4):87–91, 2006.

[MP13] Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

[MP14a] David Martens and Foster J. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–99, 2014.

[MP14b] Sanjay Modgil and Henry Prakken. The *ASPIC*$^+$ framework for structured argumentation: a tutorial. *Argument Computation*, 5(1):31–62, 2014.

[MPR$^+$19] Paras Malik, Monika Pathania, Vyas Kumar Rathaur, et al. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*, 8(7):2328–2331, 2019.

[MR20] Jean-Guy Mailly and Julien Rossit. Stability in abstract argumentation. *arXiv preprint arXiv:2012.12588*, 2020.

[OBB$^+$22] Daphne Odekerken, AnneMarie Borg, Floris Bex, Francesca Toni, Sylwia Polberg, Richard Booth, Martin Caminada, and Hiroyuki Kido. Stability and relevance in incomplete argumentation frameworks. *Frontiers in Artificial Intelligence and Applications*, 353:272 – 283, 2022.

[OBBT22] Daphne Odekerken, Floris Bex, AnneMarie Borg, and Bas Testerink. Approximating stability for applied argument-based inquiry. *Intelligent Systems with Applications*, 16:200110, 2022.

[Ove11] James Overton. Scientific explanation and computation. In Thomas Roth-Berghofer, Nava Tintarev, and David B. Leake, editors, *Proceedings of the 6th International Explanation-Aware Computing (ExaCt) Workshop*, pages 41–50, 2011.

[Pra15] Henry Prakken. *Formalising debates about law-making proposals as practical reasoning.* Springer, 2015.

[Pra18] Henry Prakken. Historical overview of formal argumentation. In Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors, *Handbook of Formal Argumentation*, pages 73–141. College Publications, 2018.

[Rai20] Arun Rai. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, 2020.

[RBDG20] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[SACP21] Ilia Stepin, José Maria Alonso, Alejandro Catalá, and Martin Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *Institute of Electrical and Electronics Engineers (IEEE) Access*, 9:11974–12001, 2021.

[TE11] Eric WK Tsang and Florian Ellsaesser. How contrastive explanation facilitates theory building. *Academy of Management Review*, 36(2):404–419, 2011.

[Tho73] Stephen Naylor Thomas. Practical reasoning in natural language. 1973.

[UW21] Markus Ulbricht and Johannes Peter Wallner. Strong explanations in abstract argumentation. In *Proceedigns of the 35th AAAI Conference on Artificial Intelligence, (AAAI)*, pages 6496–6504. AAAI Press, 2021.

[VL21a] Giulia Vilone and Luca Longo. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661, 2021.

[VL21b] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[Wal09] Douglas Walton. An overview of the use of argumentation schemes in case modeling. *Modelling Legal Cases*, pages 77–89, 2009.

[Woo06] James Woodward. Sensitive and insensitive causation. *The Philosophical Review*, 115(1):1–50, 2006.

[XC18] Yuming Xu and Claudette Cayrol. Initial sets in abstract argumentation frameworks. *Journal of Applied Non-Classical Logics*, 28(2-3):260–279, 2018.

[XUD$^+$19] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing*, volume 11839 of *Lecture Notes in Computer Science*, pages 563–574. Springer, 2019.

[YK10] Petri Ylikoski and Jaakko Kuorikoski. Dissecting explanatory power. *Philosophical studies*, 148(2):201–219, 2010.

[Yli07] Petri Ylikoski. The idea of contrastive explanandum. *Rethinking explanation*, pages 27–42, 2007.

[Zha21] Zhengwei Zhao. Analysis on the “Douyin (TikTok) mania” phenomenon based on recommendation algorithms. In *Proceedings of the international conference on new energy technology and industrial development (NETID)*, volume 235, page 3029. EDP Sciences, 2021.

[ZMLC18] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. Building more explainable artificial intelligence with argumentation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI)*, pages 8044–8046. AAAI Press, 2018.