Directionality of Attacks in Natural Language Argumentation

Marcos Cramer, Mathieu Guillaume
University of Luxembourg
marcos.cramer@uni.lu, mathieu.guillaume@uni.lu

Abstract

In formal (abstract and structured) argumentation theory, a central notion is that of an attack between a counterargument and the argument that it is challenging. Unlike the notion of an inconsistency between two statements in classical logic, this notion of an attack between arguments can be asymmetric, i.e. an argument A can attack an argument B without B attacking A. While this property of the formal systems studied by argumentation theorist has been motivated by considerations about the human practice of argumentation in natural language, there have not been any systematic studies on the connection between the directionality of attacks in argumentation-theoretic formalisms and the way humans actually interpret conflicts between arguments in a non-symmetric way. In this paper, we report on the result of two empirical cognitive studies that aim at filling this gap, one study with ordinary adults (undergraduate students) and one study with adult experts in formal argumentation theory. We interpret the results in light of the notions and distinctions defined in the ASPIC+ framework for structured argumentation, and discuss the relevance of our findings to past and future empirical studies about the link between human argumentation and formal argumentation theory.

1 Introduction

The formal study of argumentation is an important field of research within AI [Rahwan and Simari, 2009]. It consists of two major branches: In *abstract argumentation theory*, introduced by Dung [1995], one models arguments by abstracting away from their internal structure to focus on the relations of conflict between them. In *structured argumentation theory*, one additionally models the internal structure of arguments through a formal language in which arguments and counterarguments can be constructed [Besnard *et al.*, 2014]. We use the term *formal argumentation theory* to refer to both of these branches of argumentation theory together.

Two prominent frameworks for structured argumentation are the ASPIC+ framework [Modgil and Prakken, 2014] and the ABA framework [Toni, 2014]. ASPIC+ defines different kinds of attacks between arguments (rebuttal, undermining and undercut), depending on whether the counterargument challenges the conclusion of the attacked argument, a premise used by the attacked argument, or a defeasible inference rule applied in the attacked argument. Furthermore, ASPIC+ incorporates a notion of preference between arguments based on a preference relation between the defeasible premises and rules used in the arguments.

In classical logic, the inconsistency between two statements is a symmetric relation, so that it gives rise to an undirected notion of logical conflict. One of the central features of formal argumentation theory that distinguishes it from classical logic is the presence of a directed notion of logical conflict, namely unidirectional attacks. For example, in ASPIC+, an undercut (an attack on an inference rule applied in the attacked argument) is always unidirectional, while a rebuttal (an attack on the conclusion of the attacked argument) is bidirectional unless there is a strict preference of the attacking argument over the attacked argument.

It is still poorly understood to which extent the directed notion of unidirectional attacks that is present in the formalisms of argumentation theory is reflected in the way humans actually employ and evaluate natural language arguments. Two questions can be asked here: Do humans systematically interpret certain kinds of conflict in a non-symmetric, directed way? And if yes, is there any correspondence between the notion of directionality that humans employ when evaluating arguments and the criteria according to which structured argumentation frameworks like ASPIC+ and ABA determine the directionality of attacks? These questions have previously not been systematically studied. This lack of understanding puts applications of formal argumentation theory to natural language argumentation on shaky grounds and makes it difficult to assess the value of such applications.

In this paper, we report on the results of two empirical cognitive studies that address the questions mentioned above in order to contribute to an increased understanding about the nature of attacks between natural language arguments and the directionality of such attacks. The results of our study confirm that humans systematically interpret certain kinds of conflict in a directed way, and suggest that the criteria according to which ASPIC+ determines the directionality of attacks are more in line with how humans evaluate arguments than the corresponding criteria of ABA.

2 Preliminaries of Structured Argumentation

One important family of frameworks for structured argumentation is the family of ASPIC-like frameworks, consisting among others of the original ASPIC framework [Prakken, 2010], the ASPIC+ framework [Modgil and Prakken, 2014], the ASPIC$-$ framework [Caminada *et al.*, 2014] and the ASPIC-END framework [Dauphin and Cramer, 2017]. We briefly sketch ASPIC+, as it is the basis for part of our analysis of natural language argumentation.

ASPIC+ is a general framework that can be instantiated in different ways, which means that it is flexible with regards to the choice of the logical language to be used in the framework as well as the set of inference rules that are admitted. An instantiation of the ASPIC+ framework (called *argumentation theory*) is given by a formal language $\mathcal{L}$, a set of axioms over $\mathcal{L}$, a set of defeasible premises over $\mathcal{L}$, a set of strict rules and a set of defeasible rules. Arguments are built by applying the rules to deduce new information from axioms, defeasible premises or the conclusions of previous arguments. The axioms and strict rules constitute the deductive base logic underlying the argumentation theory, while the defeasible premises and rules allow for defeasible arguments to be formed, which might get rejected in the light of counterarguments.

An argument A that is part of a bigger argument B is called a *subargument* of B. Note that any argument is considered a subargument of itself. An axiom or a defeasible premise by itself also constitutes an argument, which is a subargument to any argument using this axiom or premise.

In ASPIC+, three kinds of *attacks* between arguments are distinguished: Argument A *undermines* argument B iff the conclusion of A negates a defeasible premise used in B. Argument A *rebuts* argument B iff the conclusion of A negates the conclusion of a defeasible inference made within B. A *undercuts* argument B iff the conclusion of A negates the name of a defeasible rule used in B (which intuitively means that A questions the adequacy of this defeasible rule).

Furthermore, the ASPIC+ framework allows to specify a preference ordering between the defeasible premises and rules, which gives rise to a preference order between arguments. An undermining and a rebuttal is only considered successful if the attacked argument is not preferred over the argument that attacks it.

In ASPIC+, an argument can only be accepted if all of its subarguments are accepted. For this reason, it makes sense to consider the set of all subarguments of a given argument A (including argument A itself) as a unit. This motivates the following definitions:

We say that there is a *conflict* between two arguments A and B iff some subargument of A (maybe A itself) attacks B or some subargument of B (maybe B itself) attacks A. We call a conflict between A and B *bidirectional* iff some subargument of A attacks B, and some subargument of B attacks A. A conflict that is not bidirectional is called *unidirectional*.

When there is no strict preference between A and B, a rebuttal between A and B always gives rise to a bidirectional conflict: If A rebuts B, there is always a subargument B' of B whose conclusion is negated by the conclusion of A, so that A rebuts B' and B' rebuts A. Similarly, underminings give rise to bidirectional attacks: When A undermines B by negating a premise φ used in B, then φ by itself constitutes an argument that rebuts A (or undermines A, if A is a simple argument that just states a defeasible premise). So the only way in which an attack from A to B can be unidirectional is if either A undercuts B or some subargument of A is strictly preferred to some subargument of B.

ASPIC+ is a formal framework which by itself says nothing about natural language argumentation. However, the literature on ASPIC+ is full of examples of how to use ASPIC+ to formally model the logical relationship between certain natural language arguments, and these examples were sometimes used to motivate design choices of the ASPIC+ formalism. Hence it is possible to extract predictions about the directionality of attacks between natural language arguments that are motivated by the definitions of ASPIC+. In Section 5, we will explain the predictions that we made for the studies considered in this paper, and in Section 6 we explain and discuss the results of our empirical studies concerning these predictions.

Another prominent framework for structured argumentation is the ABA *assumption-based argumentation* framework. Due to space limitations, we will not describe ABA in any detail, but just point out one important difference between ABA and ASPIC+: In ABA, an attack from A to B is always based on a direct conflict between the conclusion of A and an assumption of B. Despite some differences, assumptions in ABA work similarly to premises in ASPIC+, so that the attacks of ABA correspond roughly to the underminings in ASPIC+. But unlike in ASPIC+, these undermining-like attacks can be unidirectional even in the absence of preferences. When considering types of attacks between natural language arguments in Section 5.1, we will explain this difference between ASPIC+ and ABA with respect to the attack type *Simple Undermining*.

3 Related Work

While formal argumentation theory is an important branch of research within AI, only a few studies have empirically investigated the cognitive plausibility of the formalisms from argumentation theory. The first of its kind is the study of Rahwan *et al.* [2010], who tested how humans evaluate simple reinstatement and floating reinstatement. Their paper also includes a discussion of why this kind of empirical validation of formalisms from argumentation theory is a highly relevant method that complements the more widely applied example-based and principle-based (or postulate-based) approaches. In order to test how humans evaluate simple and floating reinstatement, they needed to formulate sets of natural language arguments that represent these two forms of reinstatement. For this, there have to be certain unidirectional attacks between the arguments, and – in the case of floating reinstatement – also a bidirectional attack. One drawback of their study is that the authors did not independently verify whether the directionality of attacks that they intended for the arguments that they designed coincide with how people interpret these arguments. This drawback is especially pressing in light of the fact that their "unidirectional" attacks directly correspond to underminings in ASPIC+, which – as explained in the previous section – actually give rise to a bidirectional conflict in ASPIC+. For this reason, we have incorporated arguments from Rahwan *et al.*'s studies in our studies, so as to test whether the assumptions they make about the directionality of attacks between natural language arguments correspond to how humans evaluate these arguments.

Among the few additional empirical cognitive studies on argumentation theory in the literature, one could noticeably refer to the following three works: Cerutti *et al.* [2014] have tested the correspondence between human evaluation of arguments and properties of a logic-programming-based approach to structured argumentation proposed by Prakken and Sartor [1997]. Rosenfeld and Kraus [2016] have empirically studied human argumentative behavior and compared it to bipolar argumentation frameworks [Cayrol and Lagasquie-Schiex, 2013]. Polberg and Hunter [2018] performed an experiment to investigate the relation between human reasoning on the one hand and bipolar and probabilistic approaches to abstract argumentation on the other hand.

Despite the importance of the directionality of attacks in the formalisms of argumentation theory, none of these studies has explicitly studied how humans evaluate the directionality of attacks between natural language arguments. The purpose of the current paper is to fill this gap.

4 Hypotheses

In classical logic, all conflicts are symmetric, i.e. an asymmetric kind of conflict like the unidirectional attacks explained in Section 2 does not exist in classical logic. While unidirectional attacks play a crucial role in most formalisms of argumentation theory, it is a priori not evident that this feature of these formalisms corresponds to a cognitively real phenomenon of human reasoning. If it does exist and has any resemblance to its formal counterparts or to the motivating examples from the argumentation-theoretic literature, it should be possible to design pairs of conflicting natural language arguments A, B whose conflict is systematically interpreted by humans in a unidirectional way. This motivates the following hypothesis:

H1. *There are conflicts between arguments that are systematically interpreted by humans as unidirectional attacks in a certain direction.*

If unidirectional attacks correspond to a cognitively real phenomenon of human reasoning, this does not necessarily mean that the criteria by which humans judge conflicts between arguments to be unidirectional rather than bidirectional correspond to the criteria put forward by the developers of frameworks of structured argumentation like ASPIC+ and ABA. Indeed, as explained in Section 2, these two frameworks do not coincide in their criteria for the directionality of attacks in the case of underminings, so that we can be sure that not both of them correspond to how humans actually reason. This motivates the following two hypotheses:

H2. *Humans evaluate the directionality of arguments according to the same criteria by which ASPIC+ determines the directionality of attacks.*

H3. *Humans evaluate underminings as unidirectional, as suggested by ABA.*

5 Design of the Studies

Before we describe the studies that we designed in order to test the three hypotheses H1, H2 and H3, we first need to explain what kinds of natural language arguments we used in these studies, and how we categorized pairs of natural language arguments depending on their *attack type*, i.e. the type of attack relation that holds between them.[1] This categorization is strongly inspired by the distinctions made in the ASPIC+ framework, but incorporates some additional distinction in order to account for properties of natural language arguments that ASPIC+ does not account for.

5.1 Attack Types for Natural Language Arguments

Rebuttal without Preference. As explained before, in ASPIC+ a rebuttal between two arguments, neither of which is preferred to the other one, is always a bidirectional attack. Our study includes pairs of natural language arguments that were designed to stand in this symmetric relation to each other: The conclusions of the two arguments directly contradict each other, and there is no information in the arguments that could justify preferring one of them over the other. We call this attack type *Rebuttal without Preference*. Here an example for this attack type:

A. A study that the Medical School of Harvard University has published in 2013 corrects mistakes made in the study by Gold et al. and concludes that only cyclic antibiotics can treat Norovirus.

B. A study that the Institute of Bacterial Sciences of Oxford University has published in 2013 corrects mistakes made in the study by Gold et al. and concludes that only non-cyclic antibiotics can treat Norovirus.

There is a contradiction between the conclusions of the two arguments due to the words *cyclic* and *non-cyclic*.

Symmetric Undermining. As explained in Section 2, we consider the sets of an argument and all its subarguments as a unit. This allows for another kind of bidirectional conflict between A and B, in which the conflict is only between a defeasible premise of A and a defeasible premise of B. In this case, the subarguments of A and B that just state these premises undermine each other. We call this bidirectional conflict between A and B *Symmetric Undermining*. Here an example for this attack type from Rahwan *et al.* [2010]:

A. Cody is a rabbit. Therefore, Cody is not a bird.

B. Cody is a cat. Therefore, Cody is not a bird.

In this example, the world knowledge that nothing can be both a rabbit and a cat is required to create a conflict. While our studies include examples of Symmetric Undermining taken over from Rahwan *et al.* [2010], most of the instances of Symmetric Undermining used in our studies are arguments that we designed ourselves.

Simple Undermining. As explained in Section 2, when a complex argument A (i.e. an argument that applies at least one rule) undermines an argument B, a subargument of B

[1] The studies involved four attack types not listed below. Due to space limitations and since we only collected small amount of data for them, we do not report about them in this paper. However, the general statements we make in this paper are all consistent with the data we have collected on these additional attack types.

rebuts A, so that the conflict between the two arguments is bidirectional in ASPIC+. But the corresponding arguments in ABA are only in unidirectional conflict, so that here ASPIC+ and ABA give rise to different predictions about natural language argumentation. We call this attack type *Simple Undermining*. In Rahwan *et al.* [2010], all arguments that were intended by the authors to be unidirectional were Simple Underminings, e.g. the following one:

A. Cody is a bird. Therefore, Cody flies.

B. Cody is a rabbit. Therefore, Cody is not a bird.

Most of the instances of Simple Undermining included in our study come from Rahwan *et al.* [2010], but we also included some instances that we designed ourselves.

Attacking an Explicit Generic. In ASPIC+, the only way to get a unidirectional conflict without preferences is through undercutting, i.e. when an argument questions the adequacy of a defeasible rule used in another argument. To design natural language analogues of this, one needs to specify what a defeasible rule is in natural language argumentation, and how a defeasible rule can be named in natural language so that it can be attacked. One idea is to consider generic statements like "Reindeer generally have antlers" as defeasible rules, and to consider a statement like "It is not right to say that reindeer generally have antlers" to be the negation of the name of this rule. However, formalizing a generic as a defeasible rule is not the only way to treat a generic in ASPIC+. It could also be formalized as a premise that contains a defeasible conditional. In this case, an attack on it would be an undermining rather than an undercutting, and would give rise to a bidirectional conflict. So we chose to call this attack type *Attacking an Explicit Generic*, so that the name of the attack type does not favor one of these interpretations over the other one. We included this attack type in our study in order to test which of these two interpretations corresponds better to how people interpret the directionality of the attack in this case. Here is an example of Attacking an Explicit Generic:

A. According to the Daily Mail, Prince William shot a reindeer yesterday. Reindeer generally have antlers. So Prince William shot an animal that has antlers.

B. The website of the International Institute for Evolutionary Biology explains that female reindeer generally don't have antlers. So it is not right to say that reindeer generally have antlers.

Undercutting Trustworthiness of Source. A further idea of how to represent undercuttings in natural language argumentation is to consider the common inference step from a statement that reports on some source S making some claim φ to the claim φ itself as the application of a defeasible rule. In this case, an argument applying this defeasible rule can be undercut by questioning the trustworthiness of source S. We call this attack type *Undercutting Trustworthiness of Source*. Here an example for this attack type:

A. The European Phonetics Centre states that the 2003 Encyclopedia of Phonetics contains many erroneous assertions and cannot be trusted.

B. The International Institute for the Advancement of Phonology states that the institution running under the name "European Phonetics Centre" is not a serious scientific institution, so its publications cannot be trusted.

Rebuttal with Preference by Specificity. A further way in which unidirectional attacks can come about in ASPIC+ is by having a strict preference relation between two arguments. ASPIC+ assumes a preference relation between defeasible premises and defeasible rules to be given, so it does not give any predictions about what kind of natural language expressions trigger a preference. However, in the literature on non-monotonic reasoning and argumentation many examples of preference have been considered, and one prominent case is that of preference by specificity, according to which more specific information has priority over more general (less specific) information. This kind of preference gives rise to the attack type *Rebuttal with Preference by Specificity*:

A. Mary put Maxy in a large cage, so Maxy cannot escape.

B. Maxy is a tiny snake, so Maxy can escape through the holes of its cage.

In this example, the information that Maxy is a pet was previously provided, and A does not add any more specific information about Maxy other than its location. B on the other hand gives very specific information about what kind of pet Maxy is, so that the rebuttal from A to B is not considered successful, and the attack is unidirectional.

Rebuttal with Preference by Recency. Another way of specifying a preference that we tested in our studies is a preference based on the recency of a scientific source of information. The idea is that statements based on more recent research are preferred over statements based on older research. This kind of preference gives rise to the attack type *Rebuttal with Preference by Recency*:

A. Specimen A consists only of amylase. The 2003 Encyclopedia of Biochemistry states that amylase is an enzyme. So specimen A consists of an enzyme.

B. A peer-reviewed research article by Smith et al. from 2006 has established that amylase is not an enzyme. Therefore no specimen consisting only of amylase consists of an enzyme.

Undercut-like Rebuttal with Preference by Recency. An additional attack type that we designed with the intension of it being a unidirectional conflict combines features of Undercutting Trustworthiness of Source and of Rebuttal with Preference by Recency. We call it *Undercut-like Rebuttal with Preference by Recency*:

A. A peer-reviewed research article by Smith et al. from 2006 has established that amylase is not an enzyme. Therefore no specimen consisting only of amylase consists of an enzyme.

B. A study that the Biology Laboratory of Harvard University has published in 2011 corrects mistakes made in the study by Smith et al. and concludes that amylase is a biologically active enzyme.

Here the conclusions of A and B are in conflict, so there is a rebuttal between them. As B is based on more recent research, it is preferred, so that the rebuttal from A to B is not successful. Additionally, the expression "corrects mistakes made in the study" questions the trustworthiness of the source of argument A as in an undercut from B to A.

No Attack. There is also the possibility of there not being any conflict whatsoever between two arguments. In order to check if our predictions about when this is the case are right, we also included such argument pairs in our study. We call this type of relation between arguments *No Attack*.

5.2 Methodology

One goal of our research was to test whether the hypotheses hold independently of the level of expertise in formal argumentation theory. Thus we conducted two empirical cognitive studies: Study 1 involved participants who were totally naive to formal argumentation theory, whereas Study 2 was conducted with experts in formal argumentation theory. Note that due to the different methodology of the two studies, we do not intend to compare their results, but just to provide two independent studies to test the hypotheses.

Study 1. Naive Adults. Twenty-seven undergraduate students from the University of Luxembourg voluntarily participated for a remuneration of 10€. They were individually tested in a quiet room. The test consisted in a questionnaire that lasted about one hour. For the test, we used the four sets of four conflicting arguments that follow the floating reinstatement structure from Rahwan *et al.* [2010], and we additionally created thirty-six sets with the same structure (i.e. corresponding to the same abstract argumentation framework) and four sets with a different structure and different number of arguments. (The list of argument sets that we created is available at `http://icr.uni.lu/mcramer/downloads/2018_Bridging_Supplement.pdf`.) We varied the nature of the attack type between conflicting arguments according to the nine categories described in previous section. Additionally, we favored variety throughout our argument sets by referring to various contexts (e.g., pet caring, hunting report, scientific publications, see previous section for illustrations).

For the test, we did not actually present any argument setr in its entirety to our participants, because we considered that judging attack relations would not be a straightforward task for non-experts. Rather, the participants were only shown two arguments at a time, and had to make a judgment about the acceptability of each of these two arguments. For this purpose, we selected 150 pairs of arguments from the designed argument sets and we divided them in three questionnaire versions of 50 items. We made sure that attack type categories were balanced across the versions. Each version was solved by a third of the participants.

Participants were instructed to judge for each argument in each pair whether they accept it, reject it, or consider its status undecided. They were explicitly told that they should not base their judgment on their knowledge, but only on the content of the arguments, and that by default an argument should be accepted, unless the other argument provides reasons to reject it. With the help of an example, it was explained to participants that non-conflicting arguments should both be accepted (no attack, denoted $A \quad B$). In another example, they were instructed to consider both arguments undecided when there is a symmetric conflict between them (in abstract argumentation theory, this corresponds to a bidirectional attack, denoted $A \leftrightarrow B$). In a last example, they were shown two arguments A and B such that B provides a reason to reject A, whereas A does not provide reason to reject B (this corresponds to a unidirectional attack from B to A, denoted $A \leftarrow B$). They were instructed to reject A and accept B in this case. In the tasks that participants had to solve, the order of the argument within the pairs was randomized across the items to avoid any effect of ordering.

Study 2. Experts in Formal Argumentation Theory. To create our sample of experts in formal argumentation theory, we contacted all authors of the chapters of a recently published scientific book about formal argumentation. All of them are thus active scientists in the field. Fourteen experts agreed to respond to our online questionnaire, which was a shortened version of the test that naive participants solved. The expert questionnaire consisted in seventeen argument sets, representative of the forty that were used in Study 1, and including some fillers with another structure to avoid any learning pattern. In this version, we presented to our expert participants the whole argument sets, not just pairs of arguments, and instructed them to indicate all the attack relations they believed existed between the arguments, by ticking the corresponding boxes. Similarly as in Study 1, the order of the arguments within the sets was also randomized to avoid ordering effect.

6 Results

Judgments made by the participants in Study 1 and Study 2 are reported (in percentage) in Table 1.

Study 1. Naive Adults. We analyzed naive adults' subjective judgment about a pair of arguments as a function of the attack type. 99.5% of responses were of one of the four forms "accept A, reject B", "reject A, accept B", "A and B undecided" or "accept A and B". By abstract argumentation theory these responses correspond to the cases "$A \rightarrow B$", "$A \leftarrow B$", "$A \leftrightarrow B$" and "$A \quad B$" respectively, which is how we report the responses in Table 1. In 0.07% of the cases the participants did not respond anything, and in 0.44% of the cases, the response was "reject A, reject B", which has no correspondence in an argumentation framework with two arguments, so we discarded theses responses. Our results revealed that the preferred attack relation differed between the attack types. Participants correctly dismissed in 83% of the cases any attack relation when there was no objective conflict between the arguments. More importantly, the majority of participants judged as bidirectional attacks argument pairs consisting of Rebuttal without Preference, Symmetric Undermining, and Simple Undermining (respectively, in 65%, 67% and 69% of the case). Conversely, they largely considered as unidirectional attacks the other conflicting situations (see Table 1). We conducted one-sample proportion tests that confirmed that all the preferred attack relations within each attack type were significantly greater than the chance level (lowest $\chi(1) = 34.241$, all $ps < .001$).

Study 2. Experts in Formal Argumentation Theory. We analyzed the judgment of experts in formal argumentation theory about the attack relation between natural language arguments as a function of the attack type. One-sample proportion tests revealed that the most largely chosen attack relation within each attack type was significantly greater than the chance level (lowest $\chi(1) = 48.931$, all $ps < .001$). Experts were excellent in dismissing non-conflicting arguments (in 98% of the cases). Importantly, the preferred attack relations were very similar to the ones chosen by naive participants (see Table 1), except for one category (Rebuttals

	ASPIC+	Study 1 (Naive)				Study 2 (Experts)			
Attack type	prediction	$A \rightarrow B$	$A \leftarrow B$	$A \leftrightarrow B$	$A \quad B$	$A \rightarrow B$	$A \leftarrow B$	$A \leftrightarrow B$	$A \quad B$
Rebuttal without Preference	$A \leftrightarrow B$	14%		**65%**	21%	0%		**86%**	14%
Symmetric Undermining	$A \leftrightarrow B$	13%		**67%**	19%	5%		**52%**	43%
Simple Undermining	$A \leftrightarrow B$	10%	16%	**69%**	4%	3%	42%	**52%**	3%
Attacking an Explicit Generic	$A \leftarrow B$	7%	**81%**	4%	7%	0%	**86%**	14%	0%
Undercutting Trustworthiness of Source	$A \leftarrow B$	11%	**77%**	4%	8%	23%	**65%**	0.4%	12%
Rebuttal with Preference by Specificity	$A \leftarrow B$	4%	**65%**	20%	11%	7%	**60%**	24%	10%
Rebuttal with Preference by Recency	$A \leftarrow B$	1%	**56%**	38%	6%	2%	36%	**57%**	5%
Undercut-like Rebuttal with Pref. by Recency	$A \leftarrow B$	7%	**83%**	4%	6%	0%	**89%**	11%	0%
No Attack	$A \quad B$	13%		3%	**83%**	2%		0%	**98%**

Table 1: Percentage of chosen attack relation in Study 1 and Study 2 as a function of the attack type. For the attack types designed with the intention to be unidirectional, $A \leftarrow B$ indicates the intended direction of attack. For the other attack types, the distinction between $A \rightarrow B$ and $A \leftarrow B$ is not meaningful, so a combined percentage is shown. The majority choice is highlighted in bold. Percentage is rounded to the nearest unit, except in the case of values less than 1%, which are rounded to the nearest tenth of a unit. Dashed horizontal lines group the attack types into three categories according to the predictions motivated by ASPIC+ (as explained in Section 5.1).

with Preference by Recency) for which experts mostly considered bidirectional attacks (in 57% of the case), while naive adults tended to judge them as unilateral attacks (in 56% of the cases).

7 Discussion of the Results

The methodologies from Study 1 and Study 2 were different, so we cannot directly and rigourously compare results from one study to the other. However, it is important to emphasise that we observed very similar tendencies (i.e., majority choices) in the two studies. Except for one attack type, Study 1 and Study 2 thus showed consistent data. Taken together, our results support H1 that it is possible to create conflicting arguments that are largely (i.e. above 80%) considered as unidirectional attacks by humans.

Hypothesis H2 by itself does not specify a correspondence between the criteria by which ASPIC+ determines the directionality of attacks between formal arguments, and analogous criteria for the directionality of attacks between natural language arguments. Since H2 can only be evaluated in light of such a correspondence, we here evaluate it based on the correspondence described in Section 5.1. If H2 is interpreted in this way, the data also confirm H2, since the majority judgment generally coincides with the ASPIC+-based predictions explained in Section 5.1, and since the majority judgment is in all cases significantly greater than the chance level.

The only discrepancy from the ASPIC+-based predictions is the expert judgment on Rebuttal with Preference by Recency, but in light of the overall trends, this is best explained by saying that most experts do not consider the recency of a scientific publication cited in an argument as a valid criterion for strictly preferring this argument. Note that of the two possible interpretations for Attacking an Explicit Generic that we provided in Section 5.1, the one that treats this attack type as a type of undercutting is confirmed by the data, while the interpretation that treats is as a type of undermining is disconfirmed, as participants interpret it significantly different to Simple Underminings (in naive, $\chi(3) = 50.637, p < .001$; in expert, $\chi(3) = 9.938, p = .019$).

Hypothesis H3 has been disconfirmed by our results: In the case of Simple Undermining, ABA predicts a unidirectional attack, while the majority judge it as a bidirectional attack. However, according to a considerable minority of judgments made by experts (42%), there is a unidirectional attack from B to A in line with the ABA prediction, which suggests that the ideas present in the ABA framework have some reflection in the way experts (or at least a significant minority of experts) judge this attack type. Note that since Rahwan *et al.* [2010] presupposed that Simple Undermining was a unidirectional attack, our findings also suggest that Rahwan *et al.*'s interpretation of their data is problematic and should be reconsidered in the light of our findings.

8 Conclusion and Outlook

Our two studies with naive and expert participants confirm our hypothesis that some conflicts between arguments are systematically interpreted by humans as unidirectional attacks. Furthermore, the studies suggest that the way the directionality between attacks is defined in ABA is problematic, while they support the definitions in ASPIC+, as long as generic statements are treated as rules that can be undercut rather than as premises that can be undermined. At the same time, the fact that depending on the attack type, humans agree with the predictions motivated by ASPIC+ to a varying degree, suggests that the distinctions made by ASPIC+ are not fine-grained enough to fully explain how humans attribute directionality to conflicts between arguments.

Since the directionality of attacks is an important feature of formalisms of argumentation theory, our research is highly relevant to the ongoing endeavor to empirically validate these formalisms. The interpretation of the results of existing studies like that of Rahwan *et al.* [2010] might have to be reconsidered in light of our findings. Future empirical studies on argumentation theory can profit both from our results and from the methodology of our studies in order to ensure that participants generally evaluate the directionality of attacks in the way intended by the designers of the study, which is a prerequisite for studying other features of formal argumentation, e.g. the correspondence between certain argumentation semantics and the way humans evaluate arguments.

References

[Besnard *et al.*, 2014] 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.

[Caminada *et al.*, 2014] Martin Caminada, Sanjay Modgil, and Nir Oren. Preferences and Unrestricted Rebut. In *Computational Models of Argument - Proceedings of COMMA 2014*, pages 209–220, 2014.

[Cayrol and Lagasquie-Schiex, 2013] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. *International Journal of Approximate Reasoning*, 54(7):876–899, 2013.

[Cerutti *et al.*, 2014] Federico Cerutti, Nava Tintarev, and Nir Oren. Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation. In Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan, editors, *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI 2014)*, pages 207–212, 2014.

[Dauphin and Cramer, 2017] Jérémie Dauphin and Marcos Cramer. ASPIC-END: Structured Argumentation with Explanations and Natural Deduction. In *Theory and Applications of Formal Argumentation (TAFA) 2017, Revised Selected Papers, LNAI 10757*, pages 51–66, 2017.

[Dung, 1995] 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):321–357, 1995.

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

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

[Prakken and Sartor, 1997] Henry Prakken and Giovanni Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7(1-2):25–75, 1997.

[Prakken, 2010] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.

[Rahwan and Simari, 2009] Iyad Rahwan and Guillermo R. Simari. *Argumentation in Artificial Intelligence*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[Rahwan *et al.*, 2010] Iyad Rahwan, Mohammed Iqbal Madakkatel, Jean-François Bonnefon, Ruqiyabi Naz Awan, and Sherief Abdallah. Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement. *Cognitive Science*, 34(8):1483–1502, 2010.

[Rosenfeld and Kraus, 2016] Ariel Rosenfeld and Sarit Kraus. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems*, 6(4):30:1–30:33, 2016.

[Toni, 2014] Francesca Toni. A tutorial on assumption-based argumentation. *Argument & Computation*, 5(1):89–117, 2014.