1

Empirical Cognitive Studies About Formal Argumentation

Federico Cerutti, Marcos Cramer, Mathieu Guillaume, Emmanuel Hadoux, Anthony Hunter, Sylwia Polberg

ABSTRACT. The evaluation of the adequacy of approaches to formal argumentation is often done through instantiations with other established formalisms, such as logic programming or non-monotonic logic. Furthermore, new developments are frequently motivated with examples of use cases that call for the additional features. While such evaluation approaches might be useful and technically sound, they often fail to show to what degree and under what circumstances they reflect human reasoning. In order to address this challenge, in recent years multiple empirical cognitive studies have been conducted to test the relationship between human behaviour and the formal models of abstract and structured argumentation. In this chapter we describe, compare and discuss these studies, taking into account their different methodological approaches. Furthermore we discuss their relevance and potential benefits for formal argumentation, and we review various open questions that are left for future research in this area.

1 Introduction

In the previous chapters of this handbook [**?**], formal argumentation has been introduced as a logical machinery for handling defeasible reasoning, citing examples that are palatable to human readers. However, this opens the question of whether formal argumentation should be the *prescriptive* model of defeasible reasoning, or a *descriptive* one. Each of these perspectives has its merits. Prescriptive formal argumentation is valuable in regulated settings, such as legal cases, political debates, or diagnosis support in medicine. However, there are various other settings that are much less regulated, or when only one of the involved parties has to adhere to certain guidelines or protocols. Common examples include patient-doctor interactions, such as persuading the patient to stop smoking or to finish a course of antibiotics. These call for more descriptive models, particularly when misusing prescriptive ones for such scenarios can be inefficient or even harmful [Nguyen and Masthoff, 2008]. Furthermore, formal argumentation could serve as a bridge between prescriptive and descriptive approaches to reasoning. For example, it has the potential to explain descriptively how prescriptive judgements on reasoning come to be accepted. And in the case of scenarios like patient-doctor interactions, formal argumentation has

the potential to produce recommendations (conceivable as *prescriptions* of a novel kind) as to which forms of arguments are known to be persuasive.

In order to ensure that the models of formal argumentation are suited for such applications to human reasoning and to bridging the gap between prescriptive and descriptive approaches, the empirically founded descriptive methodology needs to be applied to these models. This chapter describes advances that have been made in recent years in this research thread. Based on multiple studies comparing various argumentation formalisms to actual human reasoning, this chapter will describe how these formalisms perform – according to a variety of metrics – when compared to how lay people use and evaluate arguments. The ultimate research question here, first proposed in [Rahwan *et al.*, 2010], is to quantify the descriptive quality of formal argumentation mechanisms compared to human argumentation, as well as compared to human reasoning, which can be viewed as a special case of monological argumentation (see, for example, [Mercier and Sperber, 2011]).

We start our investigation by considering probably the most famous case of formal argumentation machinery, namely Dung's argumentation framework, originally proposed in [Dung, 1995], as well as its qualities in describing human reasoning in Section 2. Since its debut, Dung's framework has been extended in various ways in order to incorporate more facets of human reasoning, some with a particular aim of taking a step back from the prescriptive approach and offer a more descriptive perspective. Thus, we also look closer at these initiatives.

Since defeasible reasoning is, ultimately, about handling uncertainty in the world, it was only a matter of time that probabilistic extensions to Dung's argumentation framework were proposed [Li *et al.*, 2011; Hunter, 2012; Thimm, 2012; Hunter, 2013; Hunter and Thimm, 2017]. Section 3.1 shows the results of empirical studies using them as descriptive accounts of of human reasoning.

Dung's framework owns a part of its popularity to its simplicity as it considers arguments as atomic entities, and permits only the attack relation between them. An important subarea of argumentation expands on this by considering additional kinds of interactions that can happen between arguments. The most popular in this regard are works that distinguish between attacks and supports [Cayrol and Lagasquie-Schiex, 2013]: in Section 3.2 we review studies investigating the descriptive quality of bipolar argumentation frameworks.

Dung's argumentation frameworks consider arguments as atomic entities. However, this is not always adequate, because the arguments that humans produce can have an internal structure. The connection between the internal structure of arguments and the attack relation assumed in Dung's argumentation framework is studied by the formalisms of structured argumentation. In Section 4 we report studies that the connection of these formalisms of structured argumentation, such as ASPIC [Prakken, 2010], to human reasoning.

The descriptive approach has been applied to multiple formalisms of human reasoning, argumentation and persuasion, not just to formalisms that lie strictly

within the bounds of formal argumentation as defined by the first and the present volume of the Handbook of Formal Argumentation. While this work is outside the scope of this chapter, there are some relevant connections to the scope of this chapter. Therefore we have included an extensive discussion of related work in Section 5, namely on cognitive biases in logical reasoning tasks (Section 5.1), non-monotonic reasoning (Section 5.2), persuasion (Section 5.3), emotions (Section 5.4), argumentation schemes (Section 5.5), Bayesian argumentation (Section 5.6), and argumentation-based judgment aggregation (Section 5.7).

In Section 6, we conclude the chapter with a brief discussion of the commonalities of the studies considered in this chapters, which lie mostly in the shared methodological approach that has the potential to be developed further and become more relevant to research in formal argumentation in the future.

2 Human Reasoning and Dung Frameworks

In his seminal paper, Dung [1995] introduced abstract argumentation as a method for giving a unified account of multiple approaches in non-monotonic logic as well as of some problems from other areas. This method was explicitly motivated by the reference to how humans evaluate the acceptability of an argument based on the evaluation of all potential counterarguments. At this point, the only connection between abstract argumentation and actual human argumentation was this motivational link. But as computational argumentation established itself as a subfield of AI, more and more researchers began to apply the methodology of abstract argumentation in a way that presupposed the existence of some viable connections to actual human reasoning, e.g. for developing tools that support humans in the organization, evaluation or production of arguments, see [Cerutti *et al.*, 2018] for a survey.

This development naturally gave rise to the research question – first made explicit by Rahwan *et al.* [2010] – whether the approach of abstract argumentation has any definite connections to actual human argumentation or reasoning that could be measured through cognitive empirical studies. In the meantime multiple studies have approached this research question. Some of these works have researched the connections between actual human reasoning and the notion of *argumentation framework* as introduced by Dung [1995]. Others have explored the connections between human reasoning and some of the various extensions of Dung's original notion of argumentation frameworks. In the current section, we focus on studies of the first kind by giving an overview over their findings and over what still needs to be explored in the future. The works of the second kind will be considered in the next chapter.

2.1 Preliminaries About Dung Frameworks

In this section we define certain background notions from abstract argumentation theory as introduced by Dung [Dung, 1995] and as explained in its current state-of-the-art form by Baroni *et al.* [Baroni *et al.*, 2018].

Definition 2.1 *A* Dung framework*, also called* argumentation framework *of* AF*, is a finite directed graph $AF = \langle Ar, att \rangle$ in which the set Ar of vertices is considered to represent arguments and the set att of edges is considered to represent the attack relation between arguments, i.e. the relation between a counterargument and the argument that it attacks.*

Given an argumentation framework, we want to choose the sets of arguments for which it is rational and coherent to accept them together. A set of arguments that may be accepted together is called an *extension*. Multiple *argumentation semantics* have been defined in the literature, i.e. multiple different ways of defining extensions given an argumentation framework. Before we consider specific argumentation semantics, we first give a formal definition of the notion of *argumentation semantics*:

Definition 2.2 *An* argumentation semantics *is a function σ that maps any AF $AF = \langle Ar, att \rangle$ to a set $\sigma(AF)$ of subsets of Ar. The elements of $\sigma(AF)$ are called σ-extensions of AF.*

Remark 2.3 *We usually define an argumentation semantics σ by specifying criteria which a subset of Ar has to satisfy in order to be a σ-extension of AF.*

In this chapter we consider the *complete*, *grounded*, *preferred*, *semi-stable*, *stable*, *stage*, *CF2*, *stage2* and *SCF2 semantics*. The first five are based on the notion of *admissibility* and are therefore called *admissibility-based semantics*. The last five always choose extensions that are *naive extensions*, i.e. maximal conflict-free sets of arguments, which is why they are called *naive-based semantics*. Note that the stable semantics is the only semantics that belongs to both categories (at the price of not providing any extension at all in some scenarios). Apart from these nine semantics, we also define *naive extensions* and *SCOOC-naive extensions*, as we need them for our definition of CF2 and SCF2 semantics respectively.

Of the nine semantics defined in this section, SCF2 is the one that has been most recently introduced in the literature [Cramer and van der Torre, 2019], and is thus the only one that is not covered in [Baroni *et al.*, 2018]. Since it is the least well-known of the semantics considered in this section, we provide some intuitions about it: SCF2 semantics is based on the principle of *Strong Completeness Outside Odd Cycles*, abbreviated *SCOOC*. Informally, the SCOOC principle says that if an argument a and its attackers are not in an odd cycle, then an extension not containing any of a's attackers must contain a. The principle is based on the idea that it is generally desirable that an argument that is not attacked by any argument in a given extension should itself be in that extension. While it is possible to ensure this generally desirable property in AFs without odd cycles, this is not the case for AFs involving an odd cycle. The idea behind the SCOOC principle is to still satisfy this property as much as possible, i.e. whenever the argument under consideration and its attackers are not in an odd cycle. The SCF2 semantics is defined in a similar

way as the already well-known CF2 semantics, with the difference being that the SCOOC principle is ensured to be satisfied in each strongly connected component; Cramer and van der Torre [2019] have shown that this way the SCOOC principle turns out to be also satisfied globally.

Definition 2.4 *An att-path is a sequence* $\langle a_0, \ldots, a_n \rangle$ *of arguments where* $(a_i, a_{i+1}) \in att$ *for* $0 \leq i < n$ *and where* $a_j \neq a_k$ *for* $0 \leq j < k \leq n$ *with either* $j \neq 0$ *or* $k \neq n$*. An* odd *att*-cycle *is an att-path* $\langle a_0, \ldots, a_n \rangle$ *where* $a_0 = a_n$ *and* n *is odd.*

Definition 2.5 *Let* $AF = \langle Ar, att \rangle$ *be an AF, and let* $S \subseteq Ar$*. We write* $AF|_S$ *for the restricted AF* $\langle S, att \cap (S \times S) \rangle$*. The set* S *is called* conflict-free *iff there are no arguments* $b, c \in S$ *such that* b *attacks* c *(i.e. such that* $(b, c) \in att$*). Argument* $a \in Ar$ *is* defended *by* S *iff for every* $b \in Ar$ *such that* b *attacks* a *there exists* $c \in S$ *such that* c *attacks* b*. We define* $S^+ = \{a \in Ar \mid S \text{ attacks } a\}$ *and* $S^- = \{a \in Ar \mid a \text{ attacks some } b \in S\}$*. We define* S *to be* strongly complete outside odd cycles *iff for every argument* $a \in Ar$*, if no argument in* $\{a\} \cup \{a\}^-$ *is in an odd att-cycle and* $S \cap \{a\}^- = \emptyset$*, then* $a \in S$*.*

- S *is a* complete extension *of AF iff it is conflict-free, it defends all its arguments and it contains all the arguments it defends.*
- S *is a* stable extension *of AF iff it is conflict-free and it attacks all the arguments of* $Ar \setminus S$*.*
- S *is the* grounded extension *of AF iff it is a subset-minimal complete extension of AF.*
- S *is a* preferred extension *of AF iff it is a subset-maximal complete extension of AF.*
- S *is a* semi-stable extension *of AF iff it is a complete extension and there exists no complete extension* S_1 *such that* $S \cup S^+ \subset S_1 \cup S_1^+$*.*
- S *is a* stage extension *of AF iff* S *is a conflict-free set and there exists no conflict-free set* S_1 *such that* $S \cup S^+ \subset S_1 \cup S_1^+$*.*
- S *is a* naive extension *of AF iff* S *is a subset-maximal conflict-free set.*
- S *is a* SCOOC-naive extension *iff* S *is subset-maximal among the conflict-free subsets of* Ar *that are strongly complete outside odd cycles.*

The idea behind CF2, stage2 and SCF2 semantics is that we partition the AF into *strongly connected components* and recursively evaluate it, component by component, using a procedure called the *simplified SCC-recursive scheme*. For defining this scheme, we first need some auxiliary notions:

Definition 2.6 *Let $AF = \langle Ar, att \rangle$ be an AF, and let $a, b \in Ar$. We define $a \sim b$ iff either $a = b$ or there is an att-path from a to b and there is an att-path from b to a. The equivalence classes under the equivalence relation $\sim$ are called* strongly connected components *(SCCs) of AF. We denote the set of SCCs of AF by* SCCS_{AF}. *Given $S \subseteq Ar$, we define $D_F(S) := \{b \in Ar \mid \exists a \in S : (a, b) \in att \wedge a \not\sim b\}$.*

Definition 2.7 *Let σ be an argumentation semantics. The argumentation semantics $scc(\sigma)$ is defined as follows. Let $AF = \langle Ar, att \rangle$ be an AF, and let $S \subseteq Ar$. Then S is an $scc(\sigma)$-extension of AF iff either*

- $|\mathrm{SCCS}_{AF}| = 1$ *and S is a σ-extension of AF, or*
- $|\mathrm{SCCS}_{AF}| > 1$ *and for each $C \in \mathrm{SCCS}_{AF}$, $S \cap C$ is an $scc(\sigma)$-extension of $AF|_{C \setminus D_A F(S)}$.*

Definition 2.8 *We define* CF2, stage2 *and* SCF2 semantics *as follows:*

- CF2 semantics *is defined to be scc(naive).*
- stage2 *semantics is defined to be scc(stage).*
- *Given an AF $AF = \langle Ar, att \rangle$, a set $S \subseteq Ar$ is called a* SCF2 extension *of AF iff S is a scc(SCOOC-naive)-extension of $AF|_{Ar'}$, where $Ar' := \{a \in Ar \mid (a, a) \notin att\}$.*

Most argumentation semantics allow for the possibility of multiple extensions, so that the status of an argument depends on the choice of extension. For some purposes a single status for each argument is needed. One way to do this is through the notion of a *justification status* as defined in [Wu and Caminada, 2010] and [Baroni *et al.*, 2018], whose terminology we follow where possible. Here we focus on the *strongly accepted*, *strongly rejected* and *weakly undecided* justification statuses:

Definition 2.9 *Let $AF = \langle Ar, att \rangle$ be an AF, let σ be an argumentation semantics such that AF has at least one σ-extension, and let $a \in A$ be an argument. We say that a is* strongly accepted with respect to σ *iff for every σ-extension E of F, $a \in E$. We say that a is* strongly rejected with respect to σ *iff for every σ-extension E of F, some $b \in E$ attacks a. We say that a is* weakly undecided *iff it is neither strongly accepted nor strongly rejected.*

2.2 Empirical Cognitive Studies About Dung Frameworks

The argumentation semantics that have been proposed in the literature share some features while they differ in other respects. One feature that all major argumentation semantics have in common is the way in which they treat *simple reinstatement*, namely the fact that they all give the justification status *strongly accepted* to argument a in the AF depicted in Figure 1. One feature on which

various semantics differ is the way they treat *floating reinstatement* and *3-cycle reinstatement*, i.e. the justification status that they give to argument d in the AF depicted in Figure 2 and to argument h in the AF depicted in Figure 3: Argument d is weakly undecided with respect to the grounded semantics and the complete semantics, but is strongly accepted with respect to the other seven semantics defined above. Argument h is weakly undecided with respect to the grounded, complete, preferred and semi-stable semantics, but is strongly accepted with respect to the CF2, stage, stage2 and SCF2 semantics. (In stable semantics, the AF depicted in Figure 3 has no extension, so that the notion of a justification status cannot be meaningfully applied.)

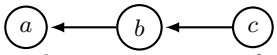


Figure 1: Simple reinstatement of argument a.

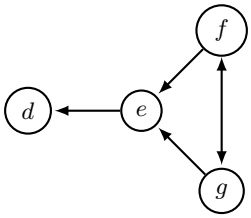


Figure 2: Floating reinstatement of argument d.

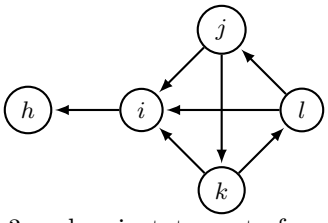


Figure 3: 3-cycle reinstatement of argument h.

This observation gives rise to two research questions with respect to the connection between actual human reasoning and abstract argumentation:

1. Do the features that all major argumentation semantics have in common (e.g. simple reinstatement) correspond to some cognitively real feature of human reasoning?

2. Which argumentation semantics can best predict human evaluation of arguments?

The first two studies that approached these research questions were performed by Rahwan *et al.* [2010]. The specific goal of their studies was to test

how humans evaluate simple reinstatement and floating reinstatement. Their two studies involved 20 and 47 participants that were randomly approached in offices and public spaces in Dubai. Participants were shown between one and four natural language arguments and asked to assess the conclusion of the highlighted argument, using a seven-point Likert scale anchored at *certainly false* and *certainly true*. The natural language arguments were designed to correspond to the arguments in the simple reinstatement or floating reinstatement AFs depicted in Figures 1 and 2 respectively. Some participants were only shown a part of those arguments. The highlighted argument that participants had to judge always corresponded to argument a or d in these two AFs.

Here is an example of an argument set used in Rahwan *et al.*'s studies as a natural language analogue of the simple reinstatement AF depicted in Figure 1:

(a) The battery of Alex's car is not working. Therefore, Alex's car will halt.

(b) The battery of Alex's car has just been changed today. Therefore, the battery of Alex's car is working.

(c) The garage was closed today. Therefore, the battery of Alex's car has not been changed today.

The results of their study suggest that confidence in the conclusion of an argument is highest in the case of an unattacked argument, lowest in the case of an argument attacked by an unattacked argument, and takes a medium value in the case of a reinstated argument, i.e. an argument that is attacked by an attacked argument. The study could find no difference between the confidence in the conclusion of a simply reinstated argument as compared to the confidence in the conclusion of a floating-reinstated argument.

Concerning the two research questions stated above, these results can be interpreted as follows:

1. The first research question gets a partially positive response. The simple reinstatement of an argument increased the confidence in the conclusion of an argument compared to the case when the argument was attacked by an unattacked argument, as is suggested by all major argumentation semantics. But this response is only partially positive, because the confidence in that conclusion did not rise back to the level of an unattacked argument, something that cannot be explained with any of the major extension-based argumentation semantics.

2. Since all standard semantics agree on the evaluation of simple reinstatement, only the results on floating reinstatement could distinguish between different semantics. Therefore only limited claims could be made concerning the second research question about which semantics best predicts human judgments. The fact that floating reinstatement is treated in the same way as simple reinstatement suggests that grounded and complete semantics are worse at predicting human behaviour than the other seven argumentation semantics defined above.

Being the first study to investigate the cognitive plausibility of the formalisms from argumentation theory, it laid the foundations for further work in this area. Cramer and Guillaume [Cramer and Guillaume, 2018b; Cramer and Guillaume, 2019] performed two further studies that expanded Rahwan *et al.*'s. One of the stated aims of these studies was to overcome some limitations of Rahwan *et al.*'s methodology.

For example, Rahwan *et al.* [2010] did not empirically test their assumption that the natural language argument sets that they designed actually correspond to the intended AFs. This limitation is especially pressing in light of the fact that the attacks that they intended to be unidirectional were based on conflicts between the conclusion of the attacking argument and the premise of the attacked argument, without any indication of a preference. In the frameworks of structured argumentation from the ASPIC family [Modgil and Prakken, 2018; Caminada *et al.*, 2014], such undermining without preferences always give rise to bidirectional attacks. An empirical study that we discuss in Section 4.4 has confirmed that humans are more likely to interpret such undermining without preferences as bidirectional attacks than as unidirectional attacks [Cramer and Guillaume, 2018a].

To overcome this limitation of Rahwan *et al.*'s study, Cramer and Guillaume first performed a study about the directionality of attacks in natural language argumentation [Cramer and Guillaume, 2018a]. Since this study compares human reasoning to predictions of structured argumentation frameworks like ASPIC+ [Modgil and Prakken, 2018], it is discussed in detail in Section 4.4 rather than here. For the current purpose, it is enough to explain that this study introduced the notion of an *attack type* between natural language arguments and discovered that some attack types are systematically interpreted as unidirectional attacks, while others are mostly interpreted as bidirectional attacks, and a third class of attack types is interpreted as a unidirectional attack by some participants and as a bidirectional attack by others participants. For their two subsequent studies on the connection between human reasoning and abstract argumentation, Cramer and Guillaume used the attack types of the first two kinds, as they had been confirmed to have a certain stable interpretation concerning the directionality of the attack relation between them.

For their first study, Cramer and Guillaume [2018b] used the findings of their prestudy on the directionality of attacks in natural language argumentation [Cramer and Guillaume, 2018a] to design the sets of three to five natural language arguments that correspond to the simple reinstatement AF, the floating reinstatement AF or the 3-cycle reinstatement AF. Including the 3-cycle reinstatement AF allowed them to distinguish some semantics that Rahwan *et al.*'s study could not distinguish. The study involved arguments based on three different thematic contexts: arguments based on news reports, arguments based on scientific publications, and arguments based on the precision of a calculation tool. As an example, here is the argument set of the scientific context corresponding to the floating reinstatement AF depicted in Figure 2:

(d) Specimen A consists only of amylase. The 1972 Encyclopaedia of Biochemistry states that amylase is an enzyme. So specimen A consists of an enzyme.

(e) A peer-reviewed research article by Smith et al. from 2006 presented new findings that amylase is not an enzyme. Therefore no specimen consisting only of amylase consists of an enzyme.

(f) 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.

(g) A study that the Biochemistry Laboratory of Oxford University has published in 2011 corrects mistakes made in the study by Smith et al. and concludes that amylase is a biologically inactive enzyme.

The study was conducted with 130 undergraduate students from the University of Luxembourg. Participants were first asked to draw the attack relation between the given arguments and then to assess the acceptability of each argument by indicating either that they *accept* the argument, that they *reject* it, or that they consider it *undecided*. The limitation to three possible responses instead of a seven-point Likert scale as in Rahwan *et al.*'s study was justified by the fact that this allows for a direct comparison of human responses with the three justification statuses of arguments that we defined at the end of Section 2.1.

For both tasks, a group discussion methodology was applied to stimulate more rational thinking: Participants first responded to the task individually, next they collaboratively discussed their responses with their peers, and finally they provided an updated individual response.

The results of this study suggest that human judgements about simple reinstatement are in line with the predictions of all major semantics, thus providing a positive response to the first research question for the case of simple reinstatement. Concerning the second research question, the study suggests that CF2, stage, stage2 and SFC2 semantics predict human behaviour best, as they were the only semantics that could predict the majority responses for all arguments in all three AFs considered in this study. Preferred and semi-stable semantics fail to predict human responses in the case of 3-cycle resinstatement, while grounded and complete semantics fail to predict human responses for both floating reinstatement and 3-cycle reinstatement. (Stable semantics is disregarded here, as it does not make a meaningful prediction for 3-cycle reinstatement.)

In a second study involving 61 undergraduate students, Cramer and Guillaume [2019] modified their methodology in order to be able to study human assessments of twelve different AFs of three to eight arguments each. For this purpose they designed a fictional scenario in which arbitrary argumentation

frameworks could be constructed in a uniform way. This allowed them to include enough different and sufficiently complex AFs to distinguish between all major argumentation semantics.

The arguments were set in the following fictional context: participants were located on an imaginary island, faced with conflicting information coming from various islanders, and they had to evaluate the arguments provided in order to hopefully find the location(s) of the buried treasure(s). All the attacks between the arguments were based on information that a certain islander is not trustworthy. As an example, here is the argument set corresponding to the floating reinstatement AF depicted in Figure 2:

(d) Islander Olivia says that there is a treasure buried near the eastern tip of the island. So we should dig up the sand near the eastern tip of the island.

(e) Islander Neil says that islander Olivia is not trustworthy and that there is a treasure buried between the two oak trees. So we should not trust what Olivia says, and we should dig up the sand between the two oak trees.

(f) Islander Lisa says that islander Mila and islander Neil are not trustworthy and that there is a treasure buried on the peak of the mountain. So we should not trust what Mila and Neil say, and we should dig up the sand on the peak of the mountain.

(g) Islander Mila says that islander Lisa and islander Neil are not trustworthy and that there is a treasure buried next to the old wall. So we should not trust what Lisa and Neil say, and we should dig up the sand next to the old wall.

In this study, the notion of an attack relation was explained in advance to participants and the intended attack relation was shown to them together with the natural language arguments, in order to ensure that participants do not overlook attacks in the case of the more complex argumentation frameworks. As in the first study, the participants assessed arguments in a three-valued way (*accept*, *reject* or *undecided*).

The results of this second study suggest that SCF2, CF2 and grounded semantics are significantly better at predicting human judgements than preferred, semi-stable, stage and stage2 semantics. The differences between SCF2, CF2 and grounded were not significant in this study. (Again, stable semantics is disregarded here, because there were multiple AFs with no stable extension.)

Note the apparent mismatch between the results of the three studies with respect to the grounded semantics. While the grounded semantics was not a good predictor of human reasoning in the first two studies, it was among the three best predictors in the third study. Cramer and Guillaume [2019] explain this apparent mismatch by pointing out that their second study used more complex argumentation frameworks, which made the reasoning task cognitively more challenging and therefore led to more participants making use of the

simplifying strategy of choosing *undecided* whenever there is some reason for doubt.[1]

2.3 Outlook on Human Reasoning and Dung Frameworks

Considering the results of the three studies in this area together, the two research questions introduced above can be partially answered as follows:

1. *Do the features that all major argumentation semantics have in common (e.g. simple reinstatement) correspond to some cognitively real feature of human reasoning?*

   This research question has only been addressed for the case of simple reinstatement, not for other features that the major argumentation semantics have in common. The existing studies suggest that the way simple reinstatement is treated by all the major argumentation semantics does indeed correspond to some cognitively real feature of human reasoning. However, the fact that a simply reinstated argument is treated in the same way as an unattacked argument might be an oversimplification.

2. *Which argumentation semantics can best predict human evaluation of arguments?*

   Taken together, the results suggest that SCF2 and CF2 semantics are the best predictors of human evaluation of arguments. For complex argumentation frameworks, the grounded semantics is also a good predictor.

Given that the responses to these two research questions are based only on three studies, all of which have some limitations in their methodology, they should be considered preliminary answers that might have to be updated by future studies.

We see two main avenues for future research in this area: On the one hand, we could attempt to design empirical studies in such a way that their findings

[1] One possible way in which this explanation hypothesized by Cramer and Guillaume could be empirically tested is by designing a study in which participants have four instead of three possible responses for each argument:

1. "There are convincing reasons for accepting the argument."
2. "There are convincing reasons for rejecting the argument."
3. "There are both reasons for accepting the argument as well as reasons for rejecting it, and the information provided is not enough to decide which of these reasons should be preferred."
4. "I don't know which of these three responses is most rational."

This way the ambiguity of the *undecided* response in Cramer and Guillaume's studies is removed, because the participants have to decide whether they are making an informed judgment about the undecidedness of the argument or whether they are unsure what the correct answer is. If Cramer and Guillaume's hypothesized explanation is correct, complex argumentation frameworks should give rise to more frequent "don't know" responses, whereas informed undecidedness (response 3) should be better predicted by SCF2 or CF2 semantics than by grounded semantics.

can be compared to the set of extensions provided by each semantics rather than just to the justification status of each argument. One way this could be done is by showing participants a set of arguments some of which are highlighted, and to ask the participants whether it would be rational to accept the highlighted arguments together while not accepting any other argument from the set. Another avenue for future research is to broaden our perspective on the first research question by considering further features that all major argumentation semantics share, e.g. that the number of attackers of an argument has no impact on the status of that argument, as long as all attackers are of the same kind (e.g. all of them are unattacked, or all of them are attacked in an equivalent way).

3 Human Reasoning and Extended Frameworks of Formal Argumentation

3.1 Empirical studies about probabilistic argumentation

Argumentation is subject to various kinds of uncertainty that can arise due to imperfections of the agents involved in a given situation, incompleteness of the available information, the types of arguments we have at hand, and much more. This can lead to, for instance, doubts concerning the structure of the graph, acceptance of arguments, or how these change when we use argumentation in dialogues and dynamic settings. One of the prominent approaches for handling such lack of confidence is probabilistic argumentation, which often provides the means of quantifying the level of uncertainty we are dealing with. The two most prominent approaches within this area are the constellations approach and the epistemic approach [Hunter, 2013], discussed in more detail in Chapter **??** of this handbook:

Constellations approach It is based on a probability distribution over the subgraphs of the argument graph ([Hunter, 2012] which extends [Dung and Thang, 2010] and [Li *et al.*, 2011]), and can be used to represent the uncertainty over the structure of the graph (i.e. whether a particular argument or attack appears in the argument graph under consideration).

Epistemic approach It is based on a probability distribution over the subsets of the arguments [Thimm, 2012; Hunter, 2013; Hunter and Thimm, 2017]. It can be used to represent the uncertainty over which arguments are believed to be accepted. The epistemic approach can be constrained (using axioms or postulates) to be consistent with Dung's semantics (see Section 2.1), but it can also be used as a potentially valuable alternative to Dung's dialectical semantics [Thimm, 2012; Hunter, 2013].

A further approach is based on labellings for arguments using *in*, *out*, and *undecided*, from [Caminada and Gabbay, 2009], augmented with *off* for denoting that the argument does not occur in the graph [Riveret and Governatori, 2016]. A probability distribution over labellings can be used to give a form of

probabilistic argumentation that overlaps with the constellations and epistemic approaches.

Hence, there are some interesting proposals for bringing probability theory into argumentation. However, empirical verification of probabilistic argumentation is an open research question and in this section we will discuss the relevant work in that area.

3.1.1 Flu Vaccine Study

In order to investigate the real-world plausibility of the constellations and epistemic approaches, Polberg and Hunter [Polberg and Hunter, 2018] undertook an exploratory study involving two dialogues concerning flu shots (one of which is presented in Table 1). The dialogues were created using statements found on the NHS and CDC websites (information sections as well as FAQ) and anti-vaccine forums, so based on information that was prepared for or widely available to the public. 40 responses were gathered per dialogue[2] using crowdsourcing techniques. This means that the participants were members of the general public, not argumentation specialists. The dialogues proceeded in steps and after each step, the participants were given the following tasks:

Agreement The participants were asked to state how much they agree or disagree with a given statement. They were allowed to choose one of the seven options (*Strongly Agree*, *Agree*, *Somewhat Agree*, *Neither Agree nor Disagree*, *Somewhat Disagree*, *Disagree*, *Strongly Disagree*) or select the answer *Don't Know*.

Explanation The participants were then asked to explain the chosen level of agreement for every statement, especially any reasons for disagreement that had not been mentioned in the dialogue.

Relation The participants were asked to state how they viewed the relation between the statements. For every listed pair, they could say whether one statement was *A good reason against*, *A somewhat good reason against*, *Somewhat related, but can't say how*, *A somewhat good reason for*, *A good reason for* the other statement or select the answer *N/A* (i.e. that the statements were unrelated).

Awareness The participants were asked which of the presented statements they had been familiar with prior to the experiment. This task was given only after the last step of the dialogue.

The answers provided by the participants were then used to analyze the argument graphs generated from the responses, whether the agreement or disagreement with the statements adhered to epistemic postulates, effect of the agreement with an argument on the relations carried out by it and vice versa, and changes in opinions on arguments occurring during the dialogue.

[2]We note that the original number of responses was much greater, but a portion of participants has been disqualified due to failed language and attention checks used in the experiment

Steps	Person	Statement	
1 to 5	P1	A	Hospital staff members do not need to receive flu shots.
1 to 5	P2	B	Hospital staff members are exposed to the flu virus a lot. Therefore, it would be good for them to receive flu shots in order to stay healthy.
2 to 5	P1	C	The virus is only airborne and it is sufficient to wear a mask in order to protect yourself. Therefore, a vaccination is not necessary.
3 to 5	P2	D	The flu virus is not just airborne, it can be transmitted through touch as well. Hence, a mask is insufficient to protect yourself against the virus.
4 to 5	P1	E	The flu vaccine causes flu in order to gain immunity. Making people sick, who otherwise might have stayed healthy, is unreasonable.
5	P2	F	The flu vaccine does not cause flu. It only has some side effects, such as headaches, that can be mistaken for flu symptoms.

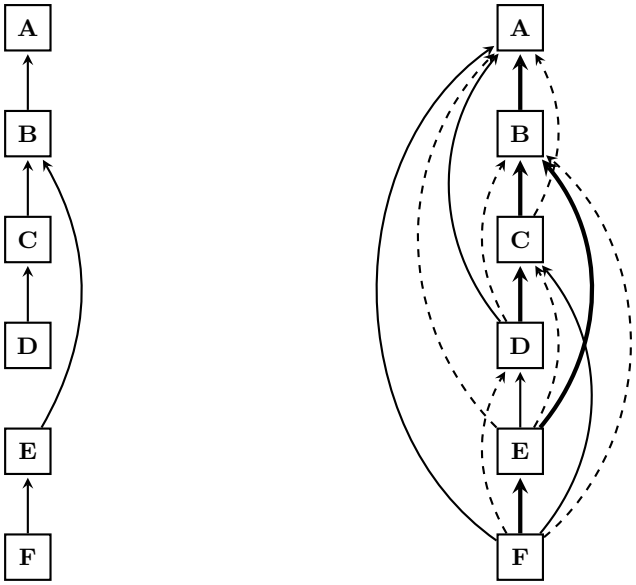
Table 1: A five-step dialogue between persons P1 and P2. This exchange starts with P1 claiming that hospital staff do not need to receive flu shots, to which P2 objects. The two counterarguments of P1 are then defeated by P2. The table presents at which steps a given statement was visible, who uttered it and what was its content.

In what follows we discuss how the findings affect the constellation and epistemic approaches.

3.1.2 Reflections on the Constellation Approach

The dialogues in the flu vaccine study proceeded in steps, by which we understand that at every step one or more arguments were added to the existing ones. At every such stage, the users stated how they viewed relations between the visible arguments. The *Relation* task answers were used to construct the graphs declared by the participants which were then compared between each other as well as to the intended graph for each dialogue stage, i.e. the graph depicting the minimal set of relations the authors have considered reasonable for a given set of arguments. Examples of such graphs can be seen in Figure 4. The similarities and disparities between all these graphs allow us to draw some conclusions for the constellation approach.

The results of the study show that in general, the most common graph (i.e. the graph that was declared by the greatest number of participants) of a given dialogue stage contained the intended graph for that stage. However, the most common graphs were not the only graph produced by the participants and there was still a significant portion of people that were of different opinions.



(a) The intended argument graph for the last step of Dialogue 1.

(b) The most commonly declared argument graph for the last step of Dialogue 1. The thicker edges represent the relations appearing in the intended graph.

Figure 4: The intended and the common graphs for the last step of Dialogue 1 from Table 1. Solid edges stand for attack and dashed for support.

As visible in Table 2, it was seen that people may interpret the statements and the relations between them differently and without adhering to the intended relations. Furthermore, as was in various cases made apparent by the answers provided in the *Explanation task*, their personal knowledge can affect their perception and evaluation of the dialogue.

There is therefore some uncertainty as to how people view relationships between arguments, and these different views can affect how they behave during a given argument exchange. The very purpose of the constellation approach is to be able to model such uncertainties concerning the topology of the argument graphs. Hence, the data from the study supports the use of the constellation approach to probabilistic argumentation for modelling the argument graphs representing the views of dialogue participants.

3.1.3 Reflections on the Epistemic Approach

The answers provided by the participants in the *Agreement* task were used to create belief distributions corresponding to these answers (e.g. *Strong agreement* response was mapped to belief of 1, representing complete belief). These were later evaluated in terms of the adherence to the epistemic postulates and

Relation	Attacking	Supporting	Dependent	N/A
(B, A)	60.50	31.50	8	0
(C, A)	29.38	63.75	5.63	1.25
(C, B)	68.75	20	7.50	3.75
(D, A)	53.33	30.83	15	0.83
(D, B)	10	84.17	5.83	0
(D, C)	66.67	25	8.33	0
(E, A)	30	45	16.25	8.75
(E, B)	53.75	16.25	21.25	8.75
(E, C)	28.75	43.75	20	7.5
(E, D)	53.75	10	27.50	8.75
(F, A)	27.50	30	40	2.5
(F, B)	10	62.50	25	2.5
(F, C)	30	27.50	37.50	5
(F, D)	0	55	37.50	7.5
(F, E)	52.50	22.50	20	5

Table 2: Occurrences of the declared relations in Dialogue 1 (values are expressed as %)

of the number of different degrees of belief the participants used.

Let us start with the epistemic postulates (see also Chapter [**?**] of this handbook). While classical semantics tend to represent a number of properties at the same time, a single postulate tends to focus on a single aspect. They therefore allow a more detailed view on the participant behaviour and can allow us to analyze the cases in which classic semantics may fail to explain it, thus providing more feedback to argumentation-based systems than classical semantics do.

The average adherence to several epistemic postulates in both dialogues can be found in Table 3. Due to their relationship with the Dung's semantics, these results offer insight into classical semantics as well. For instance, the sets of believed arguments (i.e. those with probability greater than 0.5) from rational distributions correspond to the conflict-free sets of a given argument graph. We observe that between the two dialogues, this property is generally satisfied but there is still a notable portion of participants that can at the same time accept (believe) two arguments that they perceive as conflicting. Distributions that satisfy the protective, strict, discharging and trusting postulates relate to complete extensions. In this regard, the relatively low performance of the trusting postulate (it achieved satisfaction rates of 33.5% and 43.5%) - which ensures that arguments whose attackers are all disbelieved (rejected), are believed (accepted) - highlights that participants were not eager to agree with statements just because they had no reason to disbelieve them. The performance of the discharging postulate shows that participants may disbelieve a given argument without believing any of its attackers. This is particularly important as the

study data (e.g. responses in the *Explanation* task) shows that people use their own personal knowledge in order to make judgments and might not necessarily disclose it. These results suggest that epistemic postulates can provide valuable insights into human behaviour in greater detail than Dung's dialectical semantics.

Postulate	Definition	Dialogue	
		1	2
Coherent	if for every $A, B \in \mathcal{A}$ s.t. $(A, B) \in \mathcal{R}$, $P(A) \leq 1 - P(B)$	35.5%	25%
Discharging	if for every $B \in \mathcal{A}$, if $P(B) < 0.5$ then there exists an argument $A \in \mathcal{A}$ s.t. $(A, B) \in \mathcal{R}$ and $P(A) > 0.5$	49%	60%
Founded	if $P(A) = 1$ for every initial $A \in \mathcal{A}$	23%	14%
Protective	if for every $A, B \in \mathcal{A}$ s.t. $(A, B) \in \mathcal{R}$, $P(B) > 0.5$ implies $P(A) < 0.5$	66%	49.5%
Rational	if for every $A, B \in \mathcal{A}$ s.t. $(A, B) \in \mathcal{R}$, $P(A) > 0.5$ implies $P(B) \leq 0.5$	69.5%	74%
SemiFounded	if $P(A) \geq 0.5$ for every initial $A \in \mathcal{A}$	56.5%	71.5%
SemiOptimistic	if $P(A) \geq 1 - \sum_{B \in \{A\}^-} P(B)$ for every $A \in \mathcal{A}$ that is not initial	78%	84%
Strict	if for every $A, B \in \mathcal{A}$ s.t. $(A, B) \in \mathcal{R}$, $P(A) > 0.5$ implies $P(B) < 0.5$	69.5%	61%
Trusting	if for every $B \in \mathcal{A}$, if $P(A) < 0.5$ for all $A \in \mathcal{A}$ s.t. $(A, B) \in \mathcal{R}$, then $P(B) > 0.5$	43.5%	33.5%

Table 3: Average postulate satisfaction rates by the participants in the dialogues based on the graphs obtained from the *Relation* and *Explanation* tasks. See [Polberg and Hunter, 2018] for full list of postulates considered.

Another important observation concerns the fact that the classical three-valued semantics would be insufficient to express the opinions of the majority of the participants. Most of them needed four values or more to express their beliefs and the statistical tests have shown that the choices they have made were not random. Furthermore, the changes in belief observed in the study (i.e. changes of opinions about the arguments visible between different dialogue stages) were rather subtle and observable on a fine-grained level. Not many participants changed their polarity completely (i.e. moved from agreement to disagreement or vice versa). However, some included additional clarifications in the *Explanation* task stating that while their opinions have changed for the better (or worse), they were still not sufficiently convinced to really abandon their original views.

We note that the above observation provides support for methods modelling fine-grained argument acceptability in general, not specifically just for

the epistemic approach. This observation has also been supported by a non-probabilistic study carried out by Rahwan et al [Rahwan *et al.*, 2010] that was described in Section 2.2. While it focuses on the issue of reinstatement, the results show that the level of agreement with a given argument **A** decreases once it is defeated and increases when it is defended, though still remaining significantly lower than prior to the defeat. This study, similarly to ours, lends support to the use of more fine–grained approaches towards describing the beliefs of the participants. However, as we can observe, the dialogues used in this study were much simpler and shorter than ours, and unlikely to be affected by any subjective views of the participants.

3.2 Empirical Studies About Bipolar Argumentation

One of the most prominent approaches to extending Dung's argumentation frameworks come in the form of bipolar argumentation models, which incorporate various kinds of support relations. Chapter [**?**] of this handbook contains an deep overview of these formalisms. In this section we will discuss the findings of several studies that look at support relations in the context of studies with participants.

3.2.1 Flu Vaccine Study

The study by [Polberg and Hunter, 2018] that we discussed in the previous section also produced empirical observations concerning bipolar argumentation.

In evaluating the responses from the participants, it was observed that the participants explicitly viewed certain relations as supporting. Furthermore, it was shown that the notion of defence does not account for all of the positive relations that the participants have identified between the presented statements. In particular, it was observed that there are new support relations arising in the context of the dialogue, such as support coming from statements working towards the same goal. By analysing the common graphs (i.e. the graphs declared by the highest number of participants at given steps in the dialogue - see also Section 3.1), most of the support relations that were identified by the participants can be explained as defence relations. Nevertheless, there were cases that did not fall into this category, and were more appropriately explained as support relations using bipolar argumentation. Thus, while the participants did behave in a way that was largely consistent with the notions of defence as used in dialectical semantics, they also used notions of support as conceptualized in bipolar argumentation.

It is also worth mentioning that in bipolar argumentation, mixtures of supporting and attacking links often give rise to new kinds of indirect conflicts. The study has shown that many attacks that were declared by the participants but that were not included in the intended graphs, could be reproduced by using the existing notions of indirect conflicts in these settings. Bipolar argumentation can therefore be used to model auxiliary attacks arising in the context of a dialogue, but not necessarily created on the logical level.

Another interesting observation in favour of using support relations between

arguments concerns the fact that people are not perfect reasoners. Let us consider the following two arguments uttered by opposing parties in the dialogue:

F The virus is only accompanied by stabilizers and possibly trace amounts of antibiotics used in its production.

G The vaccine contains a preservative called thimerosal which is a mercury-based compound.

The fact that the virus is accompanied only by stabilizers and antibiotics means it is not accompanied by thimerosal, which is only a preservative. This leads to a conflict between F and G. However, realizing this depends on being aware of the distinction between stabilizers and preservatives, and the participants have occasionally confused the two notions [3]. Consequently, thimerosal could have been seen as an example of a stabilizer and as a result, some participants understood G as supporting F rather than attacking it. Hence, declaring this relation differently was a conscious and somewhat justified decision, not a result of misunderstanding the exercise or an unintentional choice (a "misclick").

One's expertise and background knowledge can therefore have an impact on how relations between arguments are perceived. Furthermore, natural arguments such as the ones used here, harvested from NHS and CDC websites or general public forums, will frequently be enthymemes and rely on a given person's knowledge for correct interpretation. In such real-life situations, the use of Dung's framework can obscure the fact that various reasoning and perception issues, like the one highlighted above, are taking place. A system unable to detect that a given relation - logically intended as conflicting - is in fact seen as supporting, runs the risk of promoting undesired behaviours in the user. There is therefore a benefit in incorporating bipolar argumentation for modelling imperfect reasoners.

Thus, the vaccine dialogue study in [Polberg and Hunter, 2018] provides some evidence for the need and value of bipolar argumentation.

3.2.2 Rosenfeld and Kraus 2016 Study

In contrast to the study in [Polberg and Hunter, 2018], a less-supportive analysis of the bipolar approach can be found in [Rosenfeld and Kraus, 2016a]. This work investigated the abilities of formal argumentation, relevance heuristics, machine and transfer learning for predicting the argument choices of participants, with a particular focus on machine learning. Adequacy of computational models of argumentation was verified using three experiments. Various dialogues were sourced and then used to construct bipolar argumentation frameworks. Afterwards, the sets of arguments selected by the participants

[3] While the study was aimed at the general public and the participants did not necessarily have medical training, the statements in the study were generated based on the advice provided on websites such as NHS or CDC which are supposed to be accessible to the majority of population.

were contrasted with grounded, preferred and stable extensions of the created frameworks.

In the first experiment, consisting of 6 scenarios, the authors created bipolar argumentation frameworks which were not known to the participants, presented two standpoints from two parties and asked the participants to choose which of the additional four arguments they would use next if they were one of the participants in the discussion. In the second experiment, selected conversations from Penn Treebank Corpus were annotated and structured in the form of a bipolar argumentation frameworks. In the third experiment, a chat service was created, where participants discussed flu vaccination by using only the arguments from a predefined list. Finally, in an additional experiment, an artificial agent based on formal argumentation was implemented in order to provide suggestions to the participants during a two–person chat.

The authors report that a substantial part of the results (or in some cases, even the majority) do not conform to the outcomes predicted by the semantics. In other words, the arguments selected by the participants of the dialogues were not seen as justified. It is worth mentioning that the stated adherence to the conflict–free extension–based semantics is 78%; this is similar to the empirical results concerning the rational postulate in epistemic probabilistic argumentation, which corresponds to this semantics (see also Section 3.1.3).

Nevertheless, the causes for such behaviour of the semantics are not investigated, and the participants were not allowed to explain their decisions (the first and the third experiment) or there was no possibility to ask them for further input (the second experiment). Unlike in the study reported in [Polberg and Hunter, 2018], the participants were evaluated against the graphs constructed by the authors or annotators. As shown by the results from that study, the intended graphs do not necessarily reflect how the participants view the relations between the arguments.

There is also no discussion concerning whether these particular bipolar argument framework semantics used in this experiment [Amgoud *et al.*, 2008] are applicable. There are various ways support can be interpreted, and each of these interpretations is accompanied by several - not necessarily equivalent - types of semantics [Cayrol and Lagasquie-Schiex, 2013; Polberg and Oren, 2014; Nouioua, 2013; Gottifredi *et al.*, 2018; Cohen *et al.*, 2014; Amgoud and Ben-Naim, 2018]. The stable and conflict–free semantics used in this study [Rosenfeld and Kraus, 2016a] are based on direct and supported attacks, and only the direct ones need to be defended from and can be used for defence. This approach has been superseded by a number of different methods since it has been introduced. Consequently, the presented results indicate that these particular semantics are not useful in modelling of the user behaviour in this study, rather than there exists a deeper issue within formal argumentation itself. An additional analysis of the data, where supports are mapped to their particular interpretations and where appropriate semantics are used, would shed more light on this issue.

3.2.3 Argument Mining

Further evidence for the value of formalisms that incorporate support comes from argument mining studies that focus on obtaining arguments and relations between them from sources such as social media, Wikipedia, or Debatepedia, see for example [Cabrio and Villata, 2013; Bosc *et al.*, 2016]. These studies show that exchanges between participants often contain support relations, even more than attack relations. Similarly as in [Polberg and Hunter, 2018], these works highlight the modelling potential of bipolar argumentation and of the indirect attacks generated between arguments.

4 Empirical Studies About Structured Argumentation

Abstract argumentation, as the name suggests, abstracts away from the content of the arguments to only consider the relations between them. In contrast, structured argumentation studies how arguments can be constructed and how the relations between arguments (mostly just the attack relation) can be inferred from the structural properties of arguments. Multiple frameworks for structured argumentation have been proposed, e.g. ASPIC+ [Modgil and Prakken, 2018], ABA [Čyras *et al.*, 2017] and Prakken & Sartor System II [Prakken and Sartor, 1997]. Each of these frameworks can be instantiated in different ways by selecting strict or defeasible inference rules that correspond to some underlying logic and/or describe domain-specific inferences. Thus structured argumentation provides a bridge between logic and abstract argumentation.

Concerning the connections between structured argumentation and actual human reasoning on the basis of arguments, the following three research question can be asked:

- How can one bridge the gap between the formalisms of structured argumentation on the one hand and human arguments that are expressed in natural language, that are often presented in an enthymematic way, and whose meaning and acceptability may depend on the context, on the other hand?
- Can the existing formalisms of structured argumentation be applied to model, explain and predict the way humans construct and evaluate arguments?
- Do certain properties of particular structured argumentation formalisms correspond better to human argumentation than contrary properties of other structured argumentation formalisms (e.g. restricted vs. unrestricted rebuttal, or different ways of taking preferences into account)?

To our knowledge, there have been only three empirical studies that have compared human reasoning to frameworks of structured argumentation. Cerutti *et al.* [2014] compare human intuitions on arguments to Prakken & Sartor System II, Cramer and Guillaume [2018a] compare human judgments about arguments to ASPIC+ and ABA, and Yu *et al.* [2018] compare human judgments

about arguments to ASPIC+. So far, there has only been limited progress on the three research questions presented above. Nevertheless, we will use these three research questions as yardsticks to evaluate the contributions that these three studies have made.

4.1 Preliminaries of Structured Argumentation

Before we can look at the details of these three empirical studies, we first briefly sketch the three frameworks of structured argumentation that have been considered in these studies, namely the ASPIC+ framework [Modgil and Prakken, 2018], the ABA framework [Čyras *et al.*, 2017] and Prakken & Sartor System II [Prakken and Sartor, 1997]. We will sketch these frameworks in an informal way, focusing on the features that are relevant for the discussion of the empirical studies. For a complete formal definition of the frameworks, we refer the reader to the original works cited before.

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.

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.

The arguments and successful attacks that can be constructed on the basis of an argumentation theory give rise to an abstract argumentation framework, to which the semantics for argumentation frameworks presented in Section 2 can be applied in order to determine extensions, i.e. sets of arguments that can be coherently accepted together. A formula φ from $\mathcal{L}$ is considered *skeptically justified* with respect to the given argumentation theory iff there exists an argument with conclusion φ that is contained in every extension.

When applying ASPIC+, it is often assumed that the strict rules are *closed*

under transposition, i.e. that for any strict rule of the form $\varphi_1, \ldots, \varphi_n \to \psi$ and every $i \in \{1, \ldots, n\}$, there is a rule of the form $\bar{\psi}, \varphi_1, \ldots, \varphi_{i-1}, \varphi_{i+1}, \ldots, \varphi_n \to \bar{\varphi}_i$.

In ASPIC+ rebuttals are restricted conclusions of defeasible rules. This feature is called *restricted rebut* and has bee criticized by some authors. Caminada *et al.* [2014] propose a variant of ASPIC+ called ASPIC−, in which restricted rebut is replaced by *unrestricted rebut*, according to which an argument A attacks an argument B and any argument containing argument B if the conclusion of A negates the conclusion of B and B contains at least one defeasible rule or defeasible premise.

In *assumption-based argumentation* (*ABA*) there is only one kind of rule, which behaves like the strict rules of ASPIC+. The only source of defeasibility of arguments are therefore the defeasible premises, which in ABA are called *assumptions*. So the only way in which an argument A can attack an argument B is when the conclusion of A is the contrary of an assumption used in argument B. Due to the absence of defeasible rules in ABA, some care is needed when formalizing defeasible inferences in ABA. For example, to model the inference from Z *is an expert* and Z *said* p to p, a rule of the form

$$\mathit{expert}(Z), \mathit{said}(Z, p), \mathit{arguably}(p) \to p$$

is required, where $\mathit{arguably}(p)$ can be read as "there is no reason to doubt that p holds". One way to formally capture this intuitive meaning of $\mathit{arguably}(p)$ is by adding another rule $\neg p \to \neg \mathit{arguably}(p)$. By formalizing defeasible inferences in this way, the behaviour of ASPIC+ can be simulated within ABA, at least when preferences are not taken into account [Heyninck and Straßer, 2016].

Prakken and Sartor System II [Prakken and Sartor, 1997] is very similar to the ASPIC+ framework, with the key difference of allowing preferences to be expressed at the object level. The authors introduce an operator $\prec$ inducing binary relations between defeasible rules, and the attacks notions are thus redefined to take into consideration preferences that are the conclusions of acceptable arguments.

4.2 Prakken and Sartor and Human Intuition

Cerutti *et al.* [2014] provide evidence suggesting that when facing contradicting arguments about a course of actions, people would be comfortable being guided by a preference statement between the two options following the Prakken and Sartor System II [Prakken and Sartor, 1997]. For example, one of the pieces of text Cerutti *et al.* used is the following:

> In a TV debate, the politician AAA argues that if Region X becomes independent then X's citizens will be poorer than now. Subsequently, financial expert Dr. BBB presents a document, which scientifically shows that Region X will not be worse off financially if it becomes independent.

However, additional pieces of information, like that more recent research by several important economists that disputes the claims in the document

Dr. BBB used, can undermine the previous preference statement. This would lead people to abstain from agreeing with either one argument or the other, thus suggesting a skeptical attitude towards argumentation. In the case of the political debate, other participants were asked to assess their agreement with the politician or the financial expert (or none) based on the following, expanded, text:

> In a TV debate, the politician AAA argues that if Region X becomes independent then X's citizens will be poorer than now. Subsequently, financial expert Dr. BBB presents a document, which scientifically shows that Region X will not be worse off financially if it becomes independent. After that, the moderator of the debate reminds BBB of more recent research by several important economists that disputes the claims in that document.

Language and Context The authors of [Cerutti *et al.*, 2014] started from formal knowledge bases formalised according to Prakken and Sartor System II [Prakken and Sartor, 1997], which has the single peculiarity of allowing preferences to be expressed at the object level. They then transformed such formal knowledge bases in natural language text by handcrafting the text to represent a summary of dialogues between fictional actors in four different domains: weather forecast; political referendum about independence of a region in a country; practical argumentation towards buying a car; practical argumentation towards entering a long-term romantic relationship.

Results illustrated in [Cerutti *et al.*, 2014] suggest a correspondence between the formal theory and its representation in natural language which allows readers to reach identical conclusions. However, the authors unveil an interesting situation: in the fourth domain—looking at a decision whether entering a long-term relationship—a sort of "reversal of preference" occurs. One of the explanations the authors provide links these results to the very subjective and emotional nature of the domain. Further to this, they also candidly admit how such studies will always tussle with "collateral knowledge" [Hoffmann, 2005], or more broadly general context.

Prediction of Human Behaviour The results illustrated in [Cerutti *et al.*, 2014] suggest that people—perhaps unsurprisingly, cf. [Pinker, 2016]—possess an untaught notion of perceived logical consequence, as well as of aversion for perceived logical inconsistencies, although the authors did not explicitly give participants the option for expressing contradictory statements in their multiple-choices answers. Further, people seem more comfortable in either settling a decision—when possible—or abstaining altogether from making a judgement, thus suggesting a skeptical flavor to their innate reasoning. This seems to be consistent with the skeptical notion of acceptance provided by Prakken and Sartor System II [Prakken and Sartor, 1997].

Formalism properties and people's behaviour Looking at the results illustrated in [Cerutti *et al.*, 2014], people seem comfortable with treating preferences not as elements of a meta-language, but rather as elements of the discussion that can, in turn, be justified, undermined, or rebutted by equally

strong albeit opposite preferences. This suggests then that formalisms allowing for that might align better to people's behaviour.

4.3 Restricted vs Unrestricted Rebut

Let us imagine a dialogue between two fictional characters, where Anna tells Brenda *Jessica is a fan of two popular Korean bands, EXO and Bigbang. Both of them will hold concert series separately at nearby cities in next few weeks. So, Jessica will attend at least two concerts soon*; and Brenda replies *That won't be possible. She has been assigned too much work recently, so that she doesn't have the time to attend two concerts*?[4] According to [Yu *et al.*, 2018], agreement on Anna's stand or Brenda's stand depends on whether *restricted rebut* or *unrestricted rebut* is used.

Language and Context The authors of [Yu *et al.*, 2018] performed a study in which people were shown short, two-parties, one-round only, dialogues involving just two, potentially contradicting, arguments, similar to the ones shown at the beginning of this section. In their analysis of the study, the authors presented pairs of formal arguments in ASPIC+ or ASPIC− that can be viewed as formal analogues of those dialogues. The adequacy of the formalization was not systematically verified, but seems plausible in most cases.

Prediction of Human Behaviour Results illustrated in [Yu *et al.*, 2018] strongly suggest that people tend to agree more with an unrestricted rebut view of argumentation. This seems to indicate that people do not distinguish between defeasible arguments whose last conclusion is based on a strict rule, and defeasible arguments whose last conclusion is based on a defeasible rule. However, a caveat needs to be added to this interpretation of their results: When applying ASPIC+ strict rules can be closed under transposition, and when this is the case, the difference between restricted and unrestricted rebut disappears for the kind of conflicts that were considered in this study. So maybe the correct conclusion from the results of this study is that either people reason with unrestricted rebut or they intuitively recognize deductive inferences even when the deductive inference required for the case at hand is the transposition of a more common deductive inference.

Formalism properties and people's behaviour Based on the behaviour illustrated by participants in [Yu *et al.*, 2018], the authors call for the development of new argumentation formalisms that are both supportive of human intuition and blessed with desirable properties. Currently, using the apparently intuitive notion of unrestricted rebut clashes with the properties of *closure* and *indirect consistency* – listed among the desirable ones in [Caminada and Amgoud, 2007] – unless using the most skeptical semantics: the grounded semantics.

[4] Example formulation as presented in [Yu *et al.*, 2018].

4.4 Directionality of Attacks in Natural Language Argumentation

One of the main differences between formal argumentation theory and classical logic is that it has a directed notion of conflict, namely that of an attack from one argument to another, whereas the notion of inconsistency in classical logic is symmetric. This gives rise to the question whether there really are conflicts that humans systematically interpret as unidirectional attacks in a certain direction, and whether structured argumentation frameworks like ASPIC+ and ABA can be used as predictors for the directionality of attacks.

Cramer and Guillaume [2018a] performed two studies that addressed these questions, one study with naive participants and one with expert participants. Naive participants were shown a pair of arguments and had to determine which argument(s) they accept, which one(s) they reject and which one(s) they consider undecided. Their response could be interpreted as indication for a unidirectional attack in a certain direction, for a bidirectional attack, or for an absence of any attack. The expert study involved 14 specialists in formal argumentation that were shown sets of two to five natural language arguments and had to indicate the attack relation between them.

Language and Context Cramer and Guillaume [2018a] introduce the notion of an *attack type*. This is inspired by the distinction of three kinds of attacks in ASPIC+, but the notion of *attack type* works on natural language arguments and is more fine-grained. Examples of attack types are *Undercutting Trustworthiness of Source*, *Rebuttal with Preference by Specificity* and *Attacking an Explicit Generic* (the last one can be formalized in ASPIC+ as either an undercutting or an undermining depending on whether the attacked generic, e.g., "Reindeer generally have antlers", is formalized as a defeasible rule or as a premise that contains a defeasible rule). The studies show that some attack types are systematically interpreted as being unidirectional in a certain direction, others are mostly interpreted as being bidirectional, while a third class of attack types leads to variation between unidirectional and a bidirectional interpretation. The attack types of the first two kinds can be useful for designing empirical studies aimed at examining the relationship between abstract argumentation and human reasoning (see Section 2.2).

Prediction of Human Behaviour The studies suggest that ASPIC+ is a good predictor of the directionality of attacks, as long as generic statements are treated as rules that can be undercut rather than as premises that can be undermined. The distinction between three kinds of attacks in ASPIC+ plays an important role in this respect.

Formalism properties and people's behaviour Since ABA does not distinguish between different kinds of attacks, ABA by itself does not yield as much information as ASPIC+ that could be used to predict the directionality of attacks between natural language arguments. Cramer and Guillaume [2018a] make an even stronger claim about ABA, namely that for some attack types it makes wrong predictions about the directionality of the attack. This claim needs to be treated with care, because it depends on how defeasible inferences

are realized in ABA. Since ABA has no defeasible rules, it is the user who needs to specify how defeasible inferences are to be formalized. If they are formalized as sketched in Section 4.1, the predictions of ABA are in line with those of ASPIC+.

4.5 Outlook

Considering the results of the three studies in this area together, the three research questions introduced above can be partially answered as follows:

Concerning the first research question, one can observe that we only have a very limited understanding about how to bridge the gap between the formal approach of structured argumentation on the one hand and actual human argumentation expressed in natural language in an enthymematic and context-dependent way, on the other hand. The researchers who carried out the above studies carefully designed the natural language arguments to be used in their studies so as to minimize the impact of those aspects of human argumentation that cannot be properly captured with the existing tools of structured argumentation theory. This way they were able to make a bridge between structured argumentation and actual human argumentation, but one that cannot be easily extended to instances of human argumentation that have not been carefully designed for such studies.

Regarding the second research question, all three studies have found that human evaluation of arguments does indeed exhibit patterns that can at least partially be predicted and explained with the help of certain formalisms of structured argumentation, even if their results are so far limited to carefully designed sets of arguments. To this point, no study has attempted to use structured argumentation to predict or explain how humans construct arguments.

Concerning the third research question, it has been shown that certain features, present in some but not all structured argumentation formalisms, are indeed useful for predicting or explaining human evaluation of arguments, e.g. the possibility to argue about preferences in the object language or the possibility to distinguish different kinds of attacks. Furthermore, the results of one study suggest that human evaluation of arguments can be explained better by formalisms that either have unrestricted rebuttal or that combine restricted rebuttal with transpositions of strict rules, than by formalisms that have restricted rebuttal while lacking transpositions of strict rules.

This discussion of the existing results makes it evident that much more empirical and theoretical work is needed to provide satisfactory responses to the three research questions. Especially the first research question requires much more work that would probably require insights from natural language semantics and pragmatics as well as from more informal approaches to argumentation theory that are studied in depth by philosophers (see for example [van Eemeren *et al.*, 2014]) to be included in the design of future empirical cognitive studies on structured argumentation. A more complete response to the second research question would require to empirically study not only human evaluation of arguments but also human construction of arguments. Given the diversity

of different formalisms of structured argumentation and different variants of these formalisms, the current response to the third research question could be expanded by future studies that address the features of these formalisms not addressed so far.

5 Further Related Studies

In this section we discuss related work that is outside the scope of this chapter but has some important connections to the work presented in the rest of the chapter.

5.1 Logical Reasoning and Cognitive Biases

Cognitive psychologists generally assume that humans, from their youngest age, can reason with analogies [James *et al.*, 1890]. Humans are indeed efficient at mentally representing, manipulating, and organizing higher-order relations between mental objects. This ability is seen as one of the most crucial aspects of human cognition [Penn *et al.*, 2008], rooted in our evolutionary development, and emerging from social interactions [Tomasello, 1999]. Human reasoning had largely been associated with classical monotonic logic since psychologists from the beginning of the twentieth century predominantly considered logic as a mandatory mechanism allowing reasoning (see [Stenning and Van Lambalgen, 2012], for a review). Piaget [Piaget, 1953] for instance theorized that children progressively acquire logic (i.e., formal deductive) operations through development, and he assumed that these logic operations were mastered at adulthood. Nonetheless, empirical evidence, such as Wason's [Wason, 1968] famous observation that literate adults have severe difficulties in reasoning on abstract problems, later qualified the supposition that humans reason according to monotonic logic. On the contrary, these findings emphasized that human reasoning should be interpreted, from the monotonic logic point of view, as irrational.

Nonetheless, the irrationality explained above does not imply the existence of inherently dysfunctional cognitive structures. In this respect, the theory of mental models of reasoning [Johson-Laird and Byrne, 1991] proposes that human reasoning does not follow formal rules of inference, but alternatively depends on mental models specifically constructed for a given problematic situation. Critically, such mentally built models share their internal structure with the contextual structure of the represented real problem. Mental models are thus not abstract and are influenced by situational factors. For this reason, the way a problem is stated substantially influences the reasoning process and the decision outcome [Tversky and Kahneman, 1981; De Martino *et al.*, 2006]. Furthermore, mental models are restricted, due to physical (i.e., cognitive) limitations (see [Lenat *et al.*, 1979]). Humans cannot represent or deal with comprehensive models, and they subsequently need to build on simplified mental versions of the world. It has then been assumed that such simplified models lead to the emergence of cognitive heuristics related to the reasoning process [Shanteau, 1989; Simon, 1957; Anderson, 1986].

From a cognitive perspective, heuristics can be interpreted as mental shortcuts, used to reduce the cognitive load required by the whole reasoning process [Myers, 2010]. Critically, heuristics do not guarantee satisfactory decisions from a pure logic perspective, leading in some cases to seemingly irrational behavior, as observed by Wason [Wason, 1968]. Identifying discrepancies between human reasoning and decisions expected from monotonic logic has been the focus of many studies, and the latter emphasized the existence of sundry cognitive biases in human reasoning [Hilbert, 2012]. Cognitive biases are fundamentally and intrinsically related to human cognition, due to the heuristic nature of reasoning. Yet biases are not processing errors, they rather illustrate universal preponderating dispositions (as noted by Stanovich [Stanovich, 2003]). In this section, we briefly describe three cognitive biases that are of particular interest in formal argumentation theory.

First, it is arduous for humans to detach themselves from their perspective, because human cognition is embodied by nature [Varela *et al.*, 2016]. Thinking abstractly, without any reference to oneself or the natural world, is not an instinctive task. Consequently, to solve problems, humans are likely to elaborate on their reasoning (and make decisions) from experience or previous knowledge [Stanovich, 2003]. One famous illustration of this bias is the gambler's fallacy, where the gambler erroneously believes that a streak of a given outcome lowers the probability of observing this outcome in the future. Such a fallacy shows that people tend to evaluate the probability of a given event according to previous occurrences of similar events, although such probability does not depend on these previous occurrences [Tversky and Kahneman, 1981]. More generally, it has been showed that humans prefer to infer information outside a given problem to form an explanation – from prior knowledge – instead of using pure deductive skills from available information [Johnson-Laird *et al.*, 2004]. Human reasoning is thus biased in favor of building or manipulating mental models that are associated with the existing ones. Interestingly, this disposition can be favorable towards rational reasoning in some cases. Griggs and Cox [Griggs and Cox, 1982] indeed showed that it is possible to substantially improve performance in a difficult abstract task such as the Wason's card selection task by capitalizing on adults' experience: when provided a frame easy to relate to, humans can show great deductive skills.

Humans thus tend to reason in the light of existing mental models, so that previous knowledge or beliefs drastically influence how new information will be handled. Moreover, people tend to seek evidence in favor of their knowledge or beliefs, and they more easily accept arguments consistent with existing mental models than opposing information [Plous, 1993]. In other words, humans prefer to confirm their beliefs rather than confront them; this second propensity is a confirmatory bias. Cognitive load reduction (since the reorganization of mental models is costly) and cognitive dissonance avoidance [Festinger, 1957] are potential reasons for the existence of such bias. This confirmatory tendency implies that new information or arguments are neither neutrally processed nor

equally accepted; there is a positive bias towards decisions consistent with previous ones.

Finally, there is also a cognitive bias towards the acceptability of new arguments that are unrelated to previous mental models. In this case, humans show a truth bias, which is a predisposition to accept new information as true. This bias originates from mental model properties because they are expressed in terms of what is true (and not in terms of what is false [Johnson-Laird, 1983]). Additionally, to reduce the cognitive load, we draw conclusions as a heuristic depending on whether a conclusion holds in all, most, or some of the premises [Gilbert *et al.*, 1990; Johnson-Laird, 2006]. The criterion for acceptance is subsequently lower than the criterion for rejection. This heuristic incidentally leads to an acquiescence bias [Knowles and Nathan, 1997] in some cases, where people naturally tend to positively respond to neutral assertions. This positive truth bias notably emphasizes that there is no such neutral information in human reasoning since they convey some subjective truth.

These three (amongst many other) biases illustrate why monotonic logic should be considered irrelevant to human cognition (following Stenning & Van Lambalgen [2012]). Human reasoning is not intrinsically flawed; non-monotonic approaches of human cognition are nonetheless still needed [Ragni *et al.*, 2016]. In this respect, formal argumentation theory could bring precious insights about human reasoning. As Mercier and Sperber [Mercier and Sperber, 2011] stated, evaluating argumentation itself puts novel perspectives in the study of human irrationality. Understanding argumentation is therefore crucial to understand human cognition.

5.2 Empirical Cognitive Studies About Non-Monotonic Reasoning

From its inception in the 1990s until this day, formal argumentation has had close and fruitful interaction with the field of non-monotonic reasoning. Argumentation formalisms are often viewed as a special case of formalisms for non-monotonic reasoning. For this reason, it makes sense to compare cognitive studies about formal argumentation to cognitive studies about non-monotonic reasoning.

The field of *non-monotonic reasoning* (or *default reasoning*) started off in the late 1970s when AI researchers began to appreciate the fact that classical logic cannot account for the non-monotonic features of human reasoning, i.e. the fact that humans often retract previously drawn conclusions when learning new information. Throughout the 1980s and early 1990s various formalisms were proposed for formally capturing non-monotonic reasoning, namely quantitative approaches such as probabilistic logic, fuzzy logic and Bayesian networks as well as qualitative approaches such as default logic, circumscription, autoepistemic logic and logic programming. These formalisms and related ones continue to be studied and adapted for various purposes to this day.

Pelletier and Elio [1997; 2005] make a case for psychologism with respect to non-monotonic reasoning, i.e. for the position that the content of the field of non-monotonic reasoning is whatever reasoning patterns reside in the collective

psychological states of the population. This is, however, a highly contentious position that many practitioners in the field would either fully reject or only partially accept. For proponents of some form of psychologism with respect to non-monotonic reasoning, it certainly makes sense to empirically study how actual human reasoning compares to the various proposed formalisms of non-monotonic reasoning. But even those who reject this kind of psychologism can find value in such studies, be it because understanding human reasoning is important irrespective of whether it is the content of the field of non-monotonic reasoning, or be it because humans are to this day better at most commonsense reasoning tasks than any artificial agents and therefore understanding human reasoning better can help us build better AI tools.

In light of the large variety of non-monotonic formalisms proposed during the 1980s, Lifschitz [1988] introduced 25 *Nonmonotonic Benchmark Problems* which he argued should be modelled by every formalism that is proposed as a serious contender for modelling non-monotonic reasoning. These problems include scenarios like the following one: Given the premises listed below, it should be considered permissible to draw the conclusion listed below:

Premises:

- Blocks A and B are heavy.
- Heavy blocks are normally located on this table.
- A is not on this table.
- B is red.

Conclusion:

- B is on this table.

In many benchmark problems there is an *object-in-question* which the conclusion talks about (block B in the above example), and a different *exception-object* that according to the premises violates some default rule (block A in the above example).

Ellio and Pelletier [1993; 1996] and Pelletier and Ellio [2002] present the results of multiple empirical studies to test whether people actually draw conclusions for the aforementioned benchmark problems in line with the prescriptions proposed by Lifschitz. In Pelletier and Ellio [2005] the authors summarize and discuss the findings of these studies. Here we present a brief summary of their results followed by their interpretation.

The results of these studies by Ellio and Pelletier suggest that humans draw conclusions mostly in accordance with Lifschitz's prescriptions, and thus in accordance with the behaviour of the major qualitative non-monotonic formalisms. However, there were also some patterns in their data that could most non-monotonic formalisms cannot explain. For example, humans seem to have an inclination towards applying the following principle, called *Second-Order Default Reasoning* (or alternatively the *Guilt by Past Association* rule): "If

the available information is that the object-in-question violates other default rules, then infer that it will violate the present rule also."

Another example of a human reasoning principle that Ellio and Pelletier found in the human responses is the following principle of *Explanation-based Exceptions*: "When the given information provides both a relevant explanation of why the exception-object violates the default rule and also provides a reason to believe that the object-in-question is similar enough in this respect that it will also violate the rule, then infer that the object *does* violate the rule."

Ellio and Pelletier also compared the conclusions that humans were willing to draw based on certain information with the conclusions that humans claimed to be reasonable for a robot to draw from that same information. Here they found that people believe robots should be cautious (saying they "Can't tell") when they themselves would be willing to give a definite answer.

Another observation the authors made was that in the case of a benchmark problem in which no conclusion should be drawn according to Lifschitz's prescriptions (namely the famous *Nixon diamond* problem), half of the participants did claim that a conclusion can be drawn, but these participants were approximately equally divided between of the two potential conclusions in this problem. This might be a sign that drawing no conclusions is something that humans often try to avoid.

The experiments performed by Ellio and Pelletier involved reference to real-world categories such as various types of birds and trees that actually exist. This raises the concern that people might be using prior knowledge rather than applying only inferences based on the given premises. Hewson and Vogel (1994) and Vogel (1996) attempted to avoid this problem by formulating all premises and putative conclusions using uninterpreted Roman letters instead of English words, e.g.: "A's are normally B's." Their results suggest that people do very badly at reaching conclusions accepted in the literature. Ford and Billington (2000) point out that this bad performance might be due to the tasks being unduly meaningless when only Roman letters are used, so they propose a compromise between the two approaches: to use a fictional setting and fictional words about categories existing in this fictional setting. They performed two studies with university students and one study with academic staff who did not do research on reasoning. Their results suggest that university students do very badly at reaching conclusions accepted in the literature, whereas academic staff copes somewhat better.

Comparing their results to the results of Ellio and Pelletier's study suggests that the ability to link the provided information to existing knowledge is very important for ordinary people to be able to make reasonable conclusions, whereas academic staff is somewhat better at making reasonable conclusions even when no such link can be established.

The studies considered so far were mostly based on Lifschitz's benchmark problems and variants thereof. We will now turn our attention to studies that have aimed at determining the cognitive plausibility of the rationality

postulates for non-monotonic logic proposed by Kraus *et al.* [1990], known as the *KLM postulates*. These postulates were developed as a possible response to the question: What principles do we still accept in non-monotonic logic, once we give up the principle of monotony from classical logic? These principles are phrased in terms of the strict consequence relation $\Gamma \vdash \varphi$ (meaning that the set Γ of formulas strictly entails the formula φ) and the defeasible consequence relation $\Gamma \mid\sim \varphi$ (meaning that Γ defeasibly entails ϕ):

- Reflexivity: $\varphi \mid\sim \varphi$.
- Cut: If $\varphi \wedge \psi \mid\sim \tau$ and $\varphi \mid\sim \psi$ then $\varphi \mid\sim \tau$.
- Cautious Monotony (CM): If $\varphi \mid\sim \psi$ and $\varphi \mid\sim \tau$ then $\varphi \wedge \psi \mid\sim \tau$.
- Left Logical Equivalence (LLE): If $\varphi \vdash \psi$, $\psi \vdash \varphi$ and $\varphi \mid\sim \tau$ then $\psi \mid\sim \tau$.
- Right Weakening (RW): If $\varphi \vdash \psi$ and $\tau \mid\sim \varphi$ then $\tau \mid\sim \psi$.
- OR: If $\varphi \mid\sim \psi$ and $\tau \mid\sim \psi$, then $\varphi \vee \tau \mid\sim \psi$.

Two further related principles have received a lot of attention in the literature on non-monotonic logic:

- Rational Monotony (RM): If $\varphi \not\mid\sim \neg\psi$ and $\varphi \mid\sim \tau$ then $\varphi \wedge \psi \mid\sim \tau$.
- AND: If $\varphi \mid\sim \psi$ and $\varphi \mid\sim \tau$ then $\varphi \mid\sim \psi \wedge \tau$.

Da Silva Neves *et al.* [2002] conducted an empirical study about the cognitive plausibility of these rationality postulates of non-monotonic logic. For this purpose, they asked university students to make judgments about the degree to which different scenarios are possible. Their study involved a pre-experiment with 40 university students and a main experiment with 88 university students. In order to test the cognitive plausibility of the rationality postulates of non-monotonic logic, they performed statistical tests to determine whether the participants' judgments corroborate these postulates. Their results suggest that RW, CM, OR, AND, and RM are cognitively plausible. For the CUT rule they had different results depending on the content that was used to create concrete scenarios from the abstract patterns of the CUT rule, so that their results are not conclusive in this case. They also attempted to test the cognitive plausibility of the LLE rule, but during the test the material used turned out to be problematic in a way that did not allow them to make any conclusions about the cognitive plausibility of LLE. Moreover, their results confirmed that CM, AND and RM were validated even in cases in which Monotony does not hold.

Benferhat *et al.* [2005] tested 57 university students to test whether their reasoning is in line with various principles from non-monotonic logic. Their results suggest that on the whole, participants' reasoning was consistent with LLE, RW, OR, AND, and CUT. Concerning CM and RM their results were not conclusive.

Bonnefon *et al.* [2008] introduce a model for describing an agent's ascriptions of causality that can account for the difference between claiming that an event A causes another event B and claiming that event A facilitates event B. Their model is based on System P [Kraus *et al.*, 1990]. The authors conducted two experiments that confirmed their hypothesis that humans do actually differentiate between causality and facilitation, and broadly along the lines featured in the definitions that are built into their model.

Finally, there have also been several empirical cognitive studies that focus on probabilistic modelling of non-monotontic reasoning, e.g. [Pfeifer and Kleiter, 2005; Pfeifer and Kleiter, 2009; Pfeifer and Tulkki, 2017]. These provide further insights into the nature of non-monotonic reasong in human cognition, but perhaps are a step further away from understanding argumentation, and therefore beyond the scope of this review.

5.3 Human Reasoning and Computational Models of Persuasion

Persuasion is an activity that involves one party trying to induce another party to believe or disbelieve something, or to do (or not do) something. It is an important and complex human ability. Obviously, it is essential in commerce and politics. But, it is equally important in many aspects of daily life. Consider, for example, a child asking a parent for a raise in pocket money, a doctor trying to get a patient to enter a smoking cessation programme; a charity volunteer trying to raise funds for a poverty stricken area; or a government advisor trying to get people to avoid revealing personal details online that might be exploited by fraudsters.

Arguments are a crucial part of persuasion. They may be explicit, such as in a political debate, or they may be implicit, such as in an advert. In a dialogue involving persuasion, counterarguments also need to be taken into account. Participants may take turns in the dialogue, each of them presenting various arguments and counterarguments. So the aim of the persuader is to convince the persuadee through this exchange of arguments. Since some arguments may be more effective than others, it is valuable for the persuader to have an understanding of the persuadee and of what might work better with them.

This understanding of the persuadee can come from several, non exclusive sources: understanding their personality, their relation with the topic being discussed, or their argumentation framework concerning the ongoing persuasion process.

Most papers investigating the impact of personality on the effectiveness of persuasion are using one or other of the two most studied personality models in the psychology literature: the OCEAN model [Goldberg, 1993] (also known as the Five-Factor model or the Big Five personality traits) and the Regulatory Focus Theory [Tory Higgins, 2012]. Knowing the persuadee's relation with this model (in other words her values according to the models) allows for the prediction of her reactions to arguments and how their beliefs may evolve.

For instance, in [Lukin *et al.*, 2017], the authors used the OCEAN model to predict how the beliefs of the persuadees evolve depending on the type of

argument that was given. They created three types of arguments:

- balanced monological arguments, *i.e.*, longer pieces of text containing both viewpoints on the topic being discussed,
- emotionally-framed arguments,
- factually-framed arguments.

Interestingly, people scoring high on "Openness to Experience" (the O in the OCEAN model) were more influenced by balanced and emotional materials. On the other hand, "Agreeable" people (the A in the OCEAN model) were most affected by factual materials.

In the same line of research, the authors of [Thomas *et al.*, 2017] profiled persuadees using the OCEAN model and studied the effect of Cialdini's persuasion principles [Cialdini, 1993] on the perceived believability of arguments. They have shown, amongst other findings, that the "appeal to authority" principle is the most efficient across all personality profiles.

On the other hand, the believability of one argument can also be predicted from related arguments. In [Polberg and Hunter, 2017], the authors gathered from the participants three values associated with arguments: the believability, the convincingness and the appeal. The arguments were split in three groups depending on their source: arguments issued by "Celebrity", "Scientific" arguments and common "Society" knowledge. They have shown that, first, the believability of an argument is a good proxy for how convincing an argument is. The latter is difficult to gather, while the former is understood more easily by participants when crowdsourcing data. They have also demonstrated that people are consistent with their answers concerning arguments framed as coming from the same source, therefore showing that persuadees' profiles can be created from a small number of questions.

Personality profiles can also be used to predict high-level data on arguments. In [Hadoux and Hunter, 2019], the authors defined the notion of concerns, *i.e.*, high-level categories of arguments such as "Time"-related or "Comfort"-related arguments in the context of cycling. Using the personality profiles and demographic data as input of classification trees, they have shown that people's preferences towards the concerns can be predicted. They have also demonstrated that persuadees choose and believe arguments that are congruent with their preferences, opening up a new dimension of strategies when it comes to choosing discussion branches to enter or avoid.

Also on the role of concerns, [Chalaguine *et al.*, 2019] showed in a study with a simple chatbot, intended to persuade people to decrease meat consumption, that participants who were more concerned with environmental issues were more persuaded by positive impersonal arguments. On the other hand, participants who were more concerned with personal health were more persuaded by positive personal arguments.

Another angle to modelling the persuadees is that, instead of assuming both the persuader and the persuadee share the same argumentation framework,

the persuadee has a subset of the whole framework representing how much she knows about it. Each time something new happens, this framework evolves to include new knowledge and update the current one. In [Rosenfeld and Kraus, 2016b], the authors used a Partially Observable Markov Decision Process (POMDP) to represent the current state of the persuadee's argumentation framework and how it evolves with the interactions with the persuader. Solving the problem with POMCP [Silver and Veness, 2010], they obtained a Strategical POMDP Agent (SPA) trying to have the best estimation of the current state of the persuadee's argumentation framework and have the best strategy to maximise the persuadee's valuation of the goal argument. The experiment was about changing some students' opinion about enrolling into a Master's degree using either the SPA, a baseline or another student. The SPA and using another student were almost on par, both with a statistically significant better performance than the baseline.

5.4 Human Emotions and Computational Models of Argumentation

In addition to belief, the emotions invoked by arguments are important to take into account since they affect the way the arguments are perceived by the persuadee. Emotions are the result of how an individual appraises a stimulus. It is a cognitive process composed of a number of checks aimed at categorising a stimulus: is it relevant, what does it imply, do I have the potential to cope and is it socially significant? This process and the various patterns of checks generate different cognitive responses and coping strategies. These strategies, in turn, affect the way information is processed [Duhachek *et al.*, 2012]. For example, guilt leads to the use of active strategies focused on repairing the committed harm, whereas shame leads to the use of more passive strategies focused on the self. Combined with gain-loss framing [Tversky and Kahneman, 1981], the emotion conveyed by an argument can be used to increase the persuasiveness of this argument. While Ekman [Ekman, 1992] considered only 6 basic emotions (anger, disgust, fear, happiness, sadness and surprise), the definition and characterisation of emotions has been widely discussed in psychology. Emotions in argumentation have also been investigated recently using logic and sets of discrete emotions (see, e.g., [Nawwab *et al.*, 2010], [Lloyd-Kelly and Wyner, 2011], [Martinez *et al.*, 2012]). For instance, in [Mazzotta *et al.*, 2007], the authors analysed actual persuasion strategies and found that purely rational argumentation was rarely employed. On the other hand, emotional elements could be found everywhere. They have developed a system, PORTIA, able to create persuasive messages mixing both rational and emotional contents, using an extension of *Belief Networks* [Pearl, 1988] for the persuadee's representation.

Building upon Ekman's 6 basic emotions, the authors in [Villata *et al.*, 2017] used a combination of facial recognition and EEG to detect emotions felt by participants in a debate. The two most present emotions were anger and disgust. This is explained by the "Negative Emotion Factor" were negative emotions have a more important and lasting effect on a person's behaviour. They also

gathered the personality profiles of the participants in the frame of the OCEAN model [Goldberg, 1993]. The objective was to find correlations between certain personality traits and the strength or frequency of the emotions felt/measured. Interestingly, intuitive assumptions like "extroverted people tend to show their emotions more often" (in particular the surprise) were observed.

Taking another stance on emotions, the authors in [Hadoux *et al.*, 2018] used the "Affective Norm". It captures the emotional response to specific words in three dimensions: arousal (ranging from excited to calm), valence (pleasant to unpleasant), and dominance (from being in control to being dominated). For example, for valence scores, *leukemia* and *murder* are low and *sunshine* and *lovable* are high; for arousal scores, *grain* and *dull* are low and *lover* and *terrorism* are high; and for dominance scores, *dementia* and *earthquake* are low, and *smile* and *completion* are high. Using the database gathered by [Warriner *et al.*, 2013], containing the triplet of values for nearly 14,000 English words, they have presented a method for aggregating the values at the level of the sentence following principles from the psychology literature. They have also shown how this can be used in the context of a persuasion dialogue to calculate a strategy presenting counterarguments during a dynamic dialogue that takes generated emotions into account.

5.5 Empirical Cognitive Studies About Argumentation Schemes

Argumentation schemes represent stereotypical patterns of reasoning used in everyday conversational argumentation, and in other contexts such as legal and scientific argumentation. The schemes are accompanied by appropriate sets of critical questions which function as defeasibility conditions. An example of this is the argument from position to know [Walton *et al.*, 2008]:

Major Premise : Source a is in position to know about things in a subject domain S containing proposition A.

Minor Premise : a asserts that A is true (false).

Conclusion : A is true (false).

It can be critically questioned by raising doubts about the truth of either premise, or by asking whether a is an honest (trustworthy) source of information:

CQ1 : Is a in a position to know whether A is true (false)?

CQ2 : Is a an honest (trustworthy, reliable) source?

CQ3 : Did a assert that A is true (false)?

This section provides an overview of empirical studies related to schemes.

5.5.1 Evaluating Argumentation Schemes

In [Schellens *et al.*, 2017] participants were presented with a list of arguments and asked to rank these arguments from strongest to weakest, upon which they were asked to motivate their judgments in an interview. Such arguments were drafted as instances of five different argumentation schemes. The study confirmed that in addition to general criteria from informal logic—e.g. relevance and acceptability—people also used scheme-specific criteria, e.g. the expertise when dealing with argumentation from authority.

In [Thomas *et al.*, 2019a] participants were shown a set of five messages each promoting healthy eating, and based on different argumentation schemes. The authors then were interested in determining a reliable scale to measure the perceived persuasiveness of the arguments. The authors also show how the message types impact factors of the scale, such as effectiveness, quality, and overall perceived persuasiveness. The same authors in [Thomas *et al.*, 2019b] also proposed a tool, ArguMessage, to semi-automatically generate persuasive messages based on argumentation schemes.

In [Lazarou *et al.*, 2016] an analysis following the Cultural-Historical Activity Theory considered the teaching and learning practices in primary schools in Cyprus, showing evidence of usage of argumentation schemes.

5.5.2 Using Argumentation Schemes

In [Schneider *et al.*, 2013] the author considered a corpus of English Wikipedia deletion discussions. They also investigated the use of argumentation schemes, showing how 36% of the used argument were an instance of the Rules and Evidence schemes. In a similar type of analysis [Hansen and Walton, 2013] the authors show that the kind of argument used most frequently in the Ontario election campaign, 2011, was Appeal to Negative Consequences. Next most frequent was Practical Reasoning argumentation, followed by Appeal to Positive consequences, Argument from Sign, and Appeal to Fairness.

In [Konstantinidou and Macagno, 2013] the authors provide the evidence of the argumentative nature of students and of the benefits of using argumentation schemes as instruments for reconstructing the possible missing premises underlying their reasoning.

Another study in education, in [Song and Ferretti, 2013] 30 college students learnt two commonly used argumentation schemes (namely argument from consequences and argument from example) and critical questions associated with these schemes. Compared to the students in the contrasting conditions, those who learned critical questions wrote essays that were of higher quality and included more counterarguments, alternative standpoints, and rebuttals. In a follow-up study [Song *et al.*, 2017], the authors show that the majority of eighth-grade students they considered fail to detect fallacious arguments or clearly explain problems in the arguments they encounter. To identify whether an argument is misused or fallacious, the authors considered argumentation schemes as a golden standard. In [Green, 2015; Green, 2017] the author shows that correctly identifying the scheme an argument is an instance of is not a

trivial task even for educated professionals. As shown in [Lindahl *et al.*, 2019], even annotators with a strong background in linguistics—albeit with little explicit instructions for a given annotation task—failed to identify argumentation schemes, with the annotators agreeing neither on whole arguments nor on the units and schemes which make them up. Research in this direction is mostly looking at guidelines for the annotation of argument schemes. Musi *et al.* [2016] show that annotating argument schemes requires highly trained annotators and, in turn, an accurate annotation of both premises and claims.

On a similar note, in [Reznitskaya *et al.*, 2007] the authors provide evidence of how education in argumentation can produce benefits in reflective essays as well as in interviews. This is echoed also in [Nussbaum and Edwards, 2011] which presents a study conducted in 3 sections for 6 months (one section served as comparison group) of a 7th-grade social studies classroom in which 30 students discussed and wrote about current events adopting techniques of critical thinking closely linked to argumentation schemes and critical questions. Over time the experimental group—exposed in particular to the concepts of critical questions, and of integrative and refutational argument stratagems—made more arguments that integrated both sides of each issue. Similarly, in [Okada and Shum, 2008] the authors examine the role of Evidence-based Dialogue Maps—that exploits the Toulmin [Toulmin, 1958] argumentation scheme—as a mediating tool in scientific reasoning: as conceptual bridges for linking and making knowledge intelligible; as support for the linearisation task of generating a coherent document outline; as a reflective aid to rethinking reasoning in response to teacher feedback; and as a visual language for making arguments tangible via cartographic conventions.

5.6 Human Reasoning and Bayesian Approaches to Argumentation

Bayesian approaches to argumentation [Hahn and Hornikx, 2016] is a reaction to the MAXMIN rule for argumentation when combining linked and convergent arguments. When two or more independent arguments all support the same claim, we are in presence of *convergent arguments*. Linked arguments instead form a chain of dependencies, thus providing support for a claim only in combination.

For convergent arguments, Walton [1992] argues in favour of the MAX rule, i.e. the overall strength or plausibility of the argument is determined by the maximum of the independent arguments converging to the same claim. For linked arguments, researchers [Walton, 1992; Pollock, 2001] propose that the overall plausibility of the argument is determined by its weakest link. While some researchers [Walton, 1992] concede that there are cases where plausibility and probability are closely linked, others [Hahn *et al.*, 2013] contend that this is true in several cases. A probabilistic interpretation of the plausibility or strength of an argument leads to the conclusion that the MIN rule provides an upper bound of the probabilistic interpretation of the strength of a linked argument. Indeed, $P(A \wedge B) = P(A) \cdot P(A|B) = P(B) \cdot P(B|A) \leq$

$\min\{P(A), P(B)\}$.

5.6.1 Bayesian Argumentation

Arguments, for their defeasible nature, can be represented by a network of random variables connected in a belief network, i.e. a directed graph where nodes are random variables, and edges represent causal links. Let us consider the case of the argument from expert opinion; Walton *et al.* [2008] present a scheme for capturing it:

Major Premise : Source E is an expert in subject domain S containing proposition A.

Minor Premise : E asserts that proposition A (in domain S) is true (false).

Conclusion : A may plausibly be taken to be true (false).

Associated to the scheme there are the following six critical questions:

CQ1 : (*Expertise*) How credible is E as an expert source?

CQ2 : (*Field*) Is E an expert in the field that A is in?

CQ3 : (*Opinion*) What did E assert that implies A?

CQ4 : (*Trustworthiness*) Is E personally reliable as a source?

CQ5 : (*Consistency*) Is A consistent with what other experts assert?

CQ6 : (*Evidence*) Is E's assertion based on evidence?

However, alternative formalisations are possible, and some can make use of Bayesian inferences [Hahn *et al.*, 2013]: Figure 5 illustrates the structure of a Bayesian network[5] for representing the relevant elements of an appeal to expert opinion.

$\mathbf{X_1}$ in Figure 5 is the random variable representing whether the proposition A may plausibly be taken to be true. If A is true, then this causes the pieces of evidence ($\mathbf{X_2}$) put forward to be also true: if that is the case, then the reputation of our source of information also benefits ($\mathbf{X_3}$). However, $\mathbf{X_3}$ also depends on whether the source is trustworthy ($\mathbf{X_4}$) and a true expert of the domain ($\mathbf{X_5}$). Finally, if we consider the presence of other sources of information, such as S2 ($\mathbf{X_6}$) and S3 ($\mathbf{X_7}$) (and others if necessary), then the validity of our hypothesis $\mathbf{X_1}$ will also affect the reputation associated to them.

[5] A Bayesian network is a direct acyclic graph where nodes represent random variables, i.e. variables that can have multiple values: the easiest case is when the variable can be either true or false, but a variable can be used to represent the rolling of a dice, hence it can have six different values. Edges represent conditional and causal dependencies, e.g. smokes can cause asthma, but asthma cannot cause smokes, hence in this case there would be an arrow from a random variable representing whether an individual smokes towards a random variable representing whether the same individual suffers from asthma.

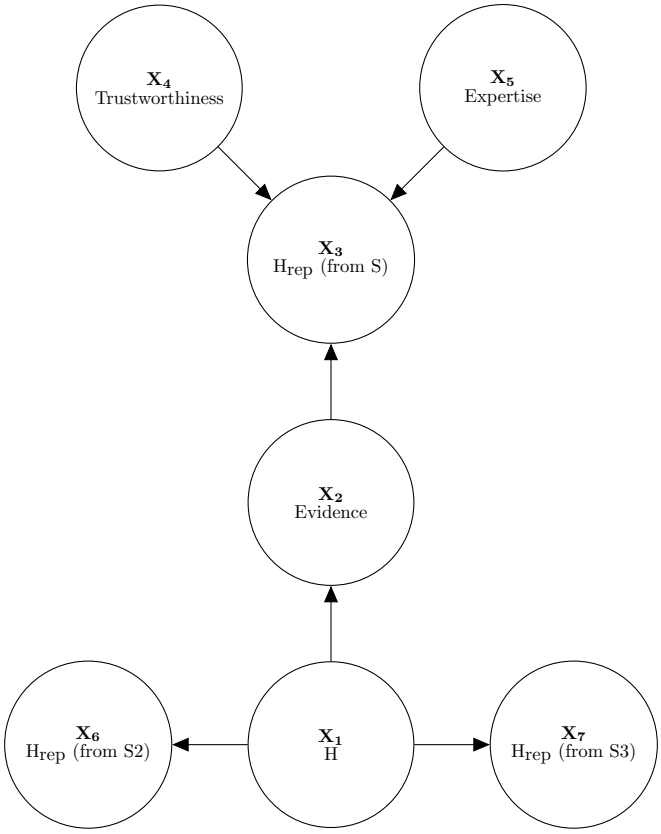


Figure 5: Structure of the Bayesian network proposed by [Hahn *et al.*, 2013] to represent the appeal to expert opinion so to be able to answer the critical questions raised by [Walton *et al.*, 2008].

Prior probability assigned to $\mathbf{X_5}$ helps answering both (*Expertise*) and (*Field*): according to [Hahn *et al.*, 2013], "the expertise will only be relevant if it is in the particular domain under consideration." In the case where what S asserts is not identical to H, then (*Opinion*) and (*Evidence*) relate to the conditional probabilities $P(\text{Evidence}|\text{H})$ and $P(\text{Evidence}|\neg\text{H})$ (or their ratio). (*Trustworthiness*) is captured by the prior assigned to $\mathbf{X_4}$, while (*Consistency*) links to variables associated to reports from different experts, i.e. $\mathbf{X_6}$ and $\mathbf{X_7}$ (and others if necessary).

5.6.2 Empirical Analyses Using Bayesian Argumentation

Bayesian argumentation provides testable measurements of the strength of the argumentation, and experimental studies such as [Oaksford and Hahn, 2004; Hahn *et al.*, 2005; Hahn and Oaksford, 2007; Corner *et al.*, 2011; Harris *et al.*, 2013] detailed how the Bayesian framework is operationalised for thus obtain-

ing qualitative and quantitative predictions and compared with lay people's perception of arguments strength. For instance, in [Oaksford and Hahn, 2004], and further in [Hahn *et al.*, 2005], the authors analysed a Bayesian account of the argument from ignorance, usually considered a reasoning fallacy. Indeed, borrowing the authors' example, the argument *Ghosts exist because no one has proved that they do not* does not seem acceptable. However, the authors' claim is that this is not because of the structure—after all, it has the same structure as *It is safe to take Ibuprofen in the recommended dose because no one has proved that it is not*—rather by the context, and thus of priors we provide to various random variables. In [Harris *et al.*, 2013], the authors also considered the *damned by faint praise* phenomenon, or *boomerang effect*, by which a very weak positive argument lead to a negative change in belief. According to the authors, this can be explained in a Bayesian framework due to an (often unstated) inference from critical missing evidence, i.e. an implicit argument from ignorance. The authors in [Hahn and Oaksford, 2007] expanded the analysis of reasoning fallacies, thus including also experiments looking at the circular arguments (*petitio principii*, and at the slippery slope argument, also expanded in [Corner *et al.*, 2011]).

5.7 Empirical Assessment of Aggregation of Argument Evaluation

Judgment aggregation is a subfield of social choice which studies how logically interrelated judgments by multiple agents can be aggregated into a group decision [List and Puppe, 2009]. Some works in judgment aggregation, e.g. Rahwan and Tohmé [2010], Caminada and Pigozzi [2011], Booth *et al.* [2014] and Awad *et al.* [2017b], have considered the problem of aggregating judgments that consist of choosing an extension of labeling of an abstract argumentation framework. Two different approaches emerged in these theoretical works: The *argument-wise plurality* rule (AWPR) chooses the collective evaluation of each argument by plurality, whereas Caminada and Pigozzi's [2011] sceptical operator, credulous operator and super credulous operator (collectively shortened as SSCOs) are based on the principle of compatibility, according to which an argument cannot be rejected if one of the agents accepted it and vice versa (but it may be accepted if some agents are undecided about it).

Awad *et al.* [2017a] performed an empirical experiment to determine which of these two approaches people consider better at aggregating opinions. For this purpose, they showed participants a set of natural language arguments that corresponded to an AF with multiple complete extensions as well as the result of a vote of members of a committee on the acceptability of the various arguments involved. Finally they were asked what decision the committee should make based on this vote. They found that AWPR was more in line with the participants' decisions than the SCSCOs, but that the difference got smaller when either the size difference between the majority and the minority in the commitment got smaller or the decision to be made by the committee was one that would personally harm an individual.

6 Conclusion and Future Work

In the field of computational argumentation, there has been an emphasis on how to represent and reason with arguments. We see this in abstract argumentation, structured argumentation, and dialogical argumentation, as well as newer topics such argument dynamics. This has involved proposals of formal systems that are then investigated in terms of theoretical properties including computational complexity, adherence to abstract postulates, forms of expressibility, etc., and the development of algorithms that are normally evaluated on randomly generated datasets.

Clearly, the research on formal argumentation has produced many interesting and potentially valuable proposals. But perhaps, the relevance to the real-world has been neglected. Related works often claim that models of argumentation more accurately reflect how humans make sense of the world, or how humans make decisions when faced with incomplete, inconsistent and uncertain information. Yet, questions about whether these formalisms actually reflect human reasoning seem to have been largely ignored by the community.

However, as this review shows, the interest in undertaking empirical studies with participants has been increasing. This has been driven by the belief that we should not just be developing theories so that they meet the intuitions of the researchers involved, but rather consider how we can use empirical evidence to inform our theoretical developments. Some of these studies are focused on whether the proposals in the literature do correctly predict or reflect human performance (e.g. studies with abstract argumentation, bipolar argumentation, and probabilistic arguments), others focus on whether the existing approaches do indeed capture various important aspects of human reasoning (e.g. whether formalisms can capture all the background knowledge that participants bring to bear on such empirical studies).

These studies offer some interesting insights, some of which support aspects of existing proposals while also suggesting that we need more sophisticated formalisms. Concerning the studies about Dung's frameworks discussed in Section 2, the studies performed so far seem to converge towards the conclusion that SCF2 and CF2 semantics are better predictors of human evaluation of arguments than other semantics studied in the literature, but more research is needed to confirm this. The other studies considered in this chapter compare human reasoning to a wide range of different formalisms from formal argumentation, so that convergence towards a common conclusion cannot be expected so far. However, this thematically disparate studies share a common methodological approach. What the studies presented in this chapter show is that this methodological approach is a fruitful addition to the methodological toolbox of formal argumentation and should be taken up and developed further in future research.

Looking forward, we would argue that the role of studies with participants needs to be expanded. There are multiple new theoretic developments in formal argumentation that are motivated by features of human argumentation

and that could benefit from an evaluation with empirical cognitive studies, for example recent advances on argument accrual [Prakken, 2019] and graded acceptibility of arguments [Grossi and Modgil, 2019]. Generally, we think that the formal argumentation community should be looking to grounding more theories with experience from such studies. Our expectation is that it may highlight some avenues for theoretical developments as more promising than others. It may help support the case for using some proposals; however, it may also flag shortcomings in current formalisms which leaves the opportunity for interesting new proposals.

BIBLIOGRAPHY

[Amgoud and Ben-Naim, 2018] Leila Amgoud and Jonathan Ben-Naim. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning*, 99:39–55, 2018.

[Amgoud *et al.*, 2008] L. Amgoud, M.C. Lagasquie-Schiex C. Cayrol, and P. Livet. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093, 2008.

[Anderson, 1986] Norman Henry Anderson. A cognitive theory of judgment and decision. In B. Brehmer, H. Jungermann, P. Lourens, and G. Sevón, editors, *New directions in research on decision making*, pages 63–108. Elsevier North-Holland, 1986.

[Awad *et al.*, 2017a] Edmond Awad, Jean-François Bonnefon, Martin Caminada, Thomas W Malone, and Iyad Rahwan. Experimental assessment of aggregation principles in argumentation-enabled collective intelligence. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–21, 2017.

[Awad *et al.*, 2017b] Edmond Awad, Richard Booth, Fernando Tohmé, and Iyad Rahwan. Judgement aggregation in multi-agent argumentation. *Journal of Logic and Computation*, 27(1):227–259, 2017.

[Baroni *et al.*, 2018] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. Abstract argumentation frameworks and their semantics. In Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors, *Handbook of Formal Argumentation*, pages 159–236. College Publications, 2018.

[Benferhat *et al.*, 2005] Salem Benferhat, Jean F Bonnefon, and Rui da Silva Neves. An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese*, 146(1-2):53–70, 2005.

[Bonnefon *et al.*, 2008] Jean-François Bonnefon, Rui Da Silva Neves, Didier Dubois, and Henri Prade. Predicting causality ascriptions from background knowledge: Model and experimental validation. *International Journal of Approximate Reasoning*, 48(3):752–765, 2008.

[Booth *et al.*, 2014] Richard Booth, Edmond Awad, and Iyad Rahwan. Interval methods for judgment aggregation in argumentation. In *Proceedings of the 14th International Conference on the Principles of Knowledge Representation and Reasoning (KR'14)*, 2014.

[Bosc *et al.*, 2016] Tom Bosc, Elena Cabrio, and Serena Villata. Tweeties squabbling: Positive and negative results in applying argument mining on social media. In Pietro Baroni, Thomas F. Gordon, Tatjana Scheffler, and Manfred Stede, editors, *Proceedings of the 6th International Conference on Computational Models of Argument (COMMA'16)*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 21–32. IOS Press, 2016.

[Cabrio and Villata, 2013] Elena Cabrio and Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument and Computation*, 4(3):209–30, 2013.

[Caminada and Amgoud, 2007] Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.

[Caminada and Gabbay, 2009] M. Caminada and D. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93:109–145, 2009.

[Caminada and Pigozzi, 2011] Martin Caminada and Gabriella Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, 22(1):64–102, 2011.

[Caminada *et al.*, 2014] Martin Caminada, Sanjay Modgil, and Nir Oren. Preferences and Unrestricted Rebut. In *Proceedings of the 5th International Conference on Computational Models of Argumentation (COMMA'14)*, 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'14)*, FAIA, pages 207–212. IOS Press, 2014.

[Cerutti *et al.*, 2018] Federico Cerutti, Sarah A. Gaggl, Matthias Thimm, and Johannes P. Wallner. Foundations of implementations for formal argumentation. In Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors, *Handbook of Formal Argumentation*, chapter 15. College Publications, February 2018. Also appears in IfCoLog Journal of Logics and their Applications 4(8):2623–2706.

[Chalaguine *et al.*, 2019] Lisa Chalaguine, Fiona Hamilton, Anthony Hunter, and Henry Potts. Impact of argument type and concerns in argumentation with a chatbot. In *Proceedings of the 31st IEEE International Conference on Tools with Artificial Intelligence (ICTAI'19)*, pages 1557–1562. IEEE, 2019.

[Cialdini, 1993] Robert B Cialdini. *Influence: The psychology of persuasion*. HarperCollins, 1993.

[Cohen *et al.*, 2014] Andrea Cohen, Sebastian Gottifredi, Alejandro Javier García, and Guillermo Ricardo Simari. A survey of different approaches to support in argumentation systems. *The Knowledge Engineering Review*, 29(5):513–550, 2014.

[Corner *et al.*, 2011] Adam Corner, Ulrike Hahn, and Mike Oaksford. The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, 64(2):133–152, feb 2011.

[Cramer and Guillaume, 2018a] Marcos Cramer and Mathieu Guillaume. Directionality of Attacks in Natural Language Argumentation. In C. Schon, editor, *Proceedings of the 4th Workshop on Bridging the Gap between Human and Automated Reasoning, co-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence {(IJCAI-ECAI} 2018)*,, volume 2261, pages 40–46. CEUR-WS.org, 2018. `http://ceur-ws.org/Vol-2261/`.

[Cramer and Guillaume, 2018b] Marcos Cramer and Mathieu Guillaume. Empirical Cognitive Study on Abstract Argumentation Semantics. *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA'18)*, pages 413–424, 2018.

[Cramer and Guillaume, 2019] Marcos Cramer and Mathieu Guillaume. Empirical Study on Human Evaluation of Complex Argumentation Frameworks. In *Proceedings of the 16th European Conference on Logics in Artificial Intelligence (JELIA'19)*, volume 11468 of *LNCS*, pages 102–115. Springer, 2019.

[Cramer and van der Torre, 2019] Marcos Cramer and Leendert van der Torre. SCF2 – an argumentation semantics for rational human judgments on argument acceptability. In Christoph Beierle, Marco Ragni, Frieder Stolzenburg, and Matthias Thimm, editors, *Proceedings of the 8th Workshop on Dynamics of Knowledge and Belief (DKB'19) and the 7th Workshop KI & Kognition (KIK'19)*, volume 2445 of *CEUR Workshop Proceedings*, pages 24–35. CEUR-WS.org, 2019.

[Čyras *et al.*, 2017] Kristijonas Čyras, X Fan, C Schulz, and F Toni. Assumption-based argumentation: disputes, explanations, preferences. *IFCoLog Journal of Logics and Their Applications*, 4(8):2407–2456, 2017.

[De Martino *et al.*, 2006] Benedetto De Martino, Dharshan Kumaran, Ben Seymour, and Raymond J Dolan. Frames, biases, and rational decision-making in the human brain. *Science*, 313(5787):684–687, 2006.

[Duhachek *et al.*, 2012] Adam Duhachek, Nidhi Agrawal, and DaHee Han. Guilt versus shame: Coping, fluency, and framing in the effectiveness of responsible drinking messages. *Journal of Marketing Research*, 49(6):928–941, 2012.

[Dung and Thang, 2010] P. M. Dung and P. M. Thang. Towards (probabilistic) argumentation for jury-based dispute resolution. In Pietro Baroni, Federico Cerutti, Massimiliano

Giacomin, and Guillermo R. Simari, editors, *Proceedings of the 3rd International Conference on Computational Models of Argumentation (COMMA'10)*, pages 171–182. IOS Press, 2010.

[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.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[Elio and Pelletier, 1993] Renée Elio and Francis Jeffry Pelletier. Human benchmarks on ai's benchmark problems. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pages 406–411, 1993.

[Elio and Pelletier, 1996] Renée Elio and Francis Jeffry Pelletier. On reasoning with default rules and exceptions. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 131–136, 1996.

[Festinger, 1957] Leon Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.

[Gilbert *et al.*, 1990] Daniel T Gilbert, Douglas S Krull, and Patrick S Malone. Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of personality and social psychology*, 59(4):601, 1990.

[Goldberg, 1993] Lewis R Goldberg. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26, 1993.

[Gottifredi *et al.*, 2018] Sebastian Gottifredi, Andrea Cohen, Alejandro Javier García, and Guillermo Ricardo Simari. Characterizing acceptability semantics of argumentation frameworks with recursive attack and support relations. *Artificial Intelligence*, 262:336–368, 2018.

[Green, 2015] Nancy Green. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argument Mining (ArgMining'15)*. Association for Computational Linguistics, 2015.

[Green, 2017] Nancy Green. Manual identification of arguments with implicit conclusions using semantic rules for argument mining. In *Proceedings of the 4th Workshop on Argument Mining (ArgMining'17)*, pages 73–78, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Griggs and Cox, 1982] Richard A Griggs and James R Cox. The elusive thematic-materials effect in wason's selection task. *British journal of psychology*, 73(3):407–420, 1982.

[Grossi and Modgil, 2019] Davide Grossi and Sanjay Modgil. On the graded acceptability of arguments in abstract and instantiated argumentation. *Artificial Intelligence*, 275:138–173, 2019.

[Hadoux and Hunter, 2019] Emmanuel Hadoux and Anthony Hunter. Comfort or Safety? Gathering and Using the Concerns of a Participant for Better Persuasion. *Argument & Computation*, Pre-press:1–35, 2019.

[Hadoux *et al.*, 2018] Emmanuel Hadoux, Anthony Hunter, and Jean-Baptiste Corrégé. *Strategic Dialogical Argumentation using Multi-Criteria Decision Making with Application to Epistemic and Emotional Aspects of Arguments Proceedings of Foiks'18, LNCS volume 10833*. pages 207-224, Springer, 2018.

[Hahn and Hornikx, 2016] Ulrike Hahn and Jos Hornikx. A normative framework for argument quality: argumentation schemes with a Bayesian foundation. *Synthese*, 193(6):1833–1873, jun 2016.

[Hahn and Oaksford, 2007] Ulrike Hahn and Mike Oaksford. The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review*, 114(3):704–732, jul 2007.

[Hahn *et al.*, 2005] Ulrike Hahn, Mike Oaksford, and Hatice Bayindir. How convinced should we be by negative evidence. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 887–892, 2005.

[Hahn *et al.*, 2013] Ulrike Hahn, Mike Oaksford, and Adam J.L. Harris. Testimony and argument: A bayesian perspective. In *Bayesian Argumentation: The Practical Side of Probability*, pages 15–38. Springer Netherlands, jan 2013.

[Hansen and Walton, 2013] Hans V. Hansen and Douglas N. Walton. Argument kinds and argument roles in the ontario provincial election, 2011. *Journal of Argumentation in Context*, 2(2):226–258, 2013.

[Harris *et al.*, 2013] Adam J. L. Harris, Adam Corner, and Ulrike Hahn. James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Thinking & Reasoning*, 19(3-4):414–429, sep 2013.

[Heyninck and Straßer, 2016] Jesse Heyninck and Christian Straßer. Relations between assumption-based approaches in nonmonotonic logic and formal argumentation. In Gabriele Kern-Isberner and Renata Wassermann, editors, *Proceedings of the 16th International Workshop on Non-Monotonic Reasoning (NMR'16)*, pages 65–76, 2016.

[Hilbert, 2012] Martin Hilbert. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin*, 138(2):211, 2012.

[Hoffmann, 2005] Michael H.G. Hoffmann. Logical argument mapping: A method for overcoming cognitive problems of conflict management. *International Journal of Conflict Management*, 16(4):304–334, 2005.

[Hunter and Thimm, 2017] A Hunter and M Thimm. Probabilistic reasoning with abstract argumentation frameworks. *Journal of Artificial Intelligence Research*, 59:565–611, 2017.

[Hunter, 2012] Anthony Hunter. Some foundations for probabilistic abstract argumentation. In *Proceedings of the 4th International Conference on Computational Models of Argumentation (COMMA'12)*, pages 117–128. IOS Press, 2012.

[Hunter, 2013] A. Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, 2013.

[James *et al.*, 1890] William James, F Burkhardt, F Bowers, and IK Skrupskelis. The principles of psychology (vol. 1, no. 2), 1890.

[Johnson-Laird *et al.*, 2004] Philip N Johnson-Laird, Vittorio Girotto, and Paolo Legrenzi. Reasoning from inconsistency to consistency. *Psychological Review*, 111(3):640, 2004.

[Johnson-Laird, 1983] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

[Johnson-Laird, 2006] Philip Nicholas Johnson-Laird. *How we reason*. Oxford University Press, USA, 2006.

[Johson-Laird and Byrne, 1991] Philip N Johson-Laird and Ruth MJ Byrne. *Deduction*. Lawrence Erlbaum Associates, 1991.

[Knowles and Nathan, 1997] Eric S Knowles and Kobi T Nathan. Acquiescent responding in self-reports: cognitive style or social concern? *Journal of Research in Personality*, 31(2):293–301, 1997.

[Konstantinidou and Macagno, 2013] Aikaterini Konstantinidou and Fabrizio Macagno. Understanding students' reasoning: Argumentation schemes as an interpretation method in science education. *Science & Education*, 22(5):1069–1087, May 2013.

[Kraus *et al.*, 1990] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.

[Lazarou *et al.*, 2016] Demetris Lazarou, Rosamund Sutherland, and Sibel Erduran. Argumentation in science education as a systemic activity: An activity-theoretical perspective. *International Journal of Educational Research*, 79:150 – 166, 2016.

[Lenat *et al.*, 1979] Douglas B Lenat, Frederick Hayes-Roth, and Philip Klahr. Cognitive economy in artificial intelligence systems. In *Proceedings of the 6th International Joint Conference on Artificial intelligence (IJCAI'79)*, pages 531–536, 1979.

[Li *et al.*, 2011] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *Proceedings of the 1st International Workshop on the Theory and Applications of Formal Argumentation (TAFA'11)*, 2011.

[Lifschitz, 1988] V Lifschitz. Benchmark problems in nonmonotonic reasoning. In *Proceedings of the 2nd International Workshop on Non-Monotonic Reasoning (NMR'88)*, volume 346 of *LNCS*, pages 202–219. Springer, 1988.

[Lindahl *et al.*, 2019] Anna Lindahl, Lars Borin, and Jacobo Rouces. Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining (ArgMining'19)*. Association for Computational Linguistics, 2019.

[List and Puppe, 2009] Christian List and Clemens Puppe. Judgment aggregation: A survey. In *Handbook of Rational and Social Choice*. Oxford University Press, Oxford, 2009.

[Lloyd-Kelly and Wyner, 2011] Martyn Lloyd-Kelly and Adam Wyner. Arguing about emotion. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*, pages 355–367. Springer, 2011.

[Lukin *et al.*, 2017] S. Lukin, P. Anand, M. Walker, and S. Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 742–753. ACL, 2017.

[Martinez *et al.*, 2012] DC Martinez, GR Simari, et al. Emotion-directed argument awareness for autonomous agent reasoning. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 15(50):30–45, 2012.

[Mazzotta *et al.*, 2007] I. Mazzotta, F. de Rosis, and V. Carofiglio. Portia: A user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent Systems*, 22(6):42–51, Nov 2007.

[Mercier and Sperber, 2011] Hugo Mercier and Dan Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(02):57–74, April 2011.

[Modgil and Prakken, 2018] Sanjay Modgil and H Prakken. Abstract rule-based argumentation. In *Handbook of Formal Argumentation*, volume 1, pages 287–364. College Publications, 2018.

[Musi *et al.*, 2016] Elena Musi, Debanjan Ghosh, and Smaranda Muresan. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the 3rd Workshop on Argument Mining (ArgMining'16)*, pages 82–93, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Myers, 2010] David G Myers. *Social psychology*. McGraw-Hill, tenth edition, 2010.

[Nawwab *et al.*, 2010] Fahd Saud Nawwab, Paul E Dunne, and Trevor JM Bench-Capon. Exploring the role of emotions in rational decision making. In Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Guillermo R. Simari, editors, *Proceedings of the 3rd International Conference on Computational Models of Argumentation (COMMA'10)*, pages 367–378, 2010.

[Neves *et al.*, 2002] Rui Da Silva Neves, Jean-François Bonnefon, and Eric Raufaste. An empirical test of patterns for nonmonotonic inference. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):107–130, 2002.

[Nguyen and Masthoff, 2008] Hien Nguyen and Judith Masthoff. Designing Persuasive Dialogue Systems: Using Argumentation with Care. In Harri Oinas-Kukkonen, Per F. V. Hasle, Marja Harjumaa, Katarina Segerståhl, and Peter Øhrstrøm, editors, *Proceedings of the 3rd International Conference on Persuasive Technology (PERSUASIVE'08)*, volume 5033 of *LNCS*, pages 201–212. Springer, 2008.

[Nouioua, 2013] Farid Nouioua. Afs with necessities: Further semantics and labelling characterization. In Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, editors, *Proceedings of the 7th International Conference on Scalable Uncertainty Management (SUM'13)*, pages 120–133. Springer, 2013.

[Nussbaum and Edwards, 2011] E. Michael Nussbaum and Ordene V. Edwards. Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences*, 20(3):443–488, 2011.

[Oaksford and Hahn, 2004] Mike Oaksford and Ulrike Hahn. A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58(2):75–85, 2004.

[Okada and Shum, 2008] Alexandra Okada and Simon Buckingham Shum. Evidence-based dialogue maps as a research tool to investigate the quality of school pupils' scientific argumentation. *International Journal of Research & Method in Education*, 31(3):291–315, 2008.

[Pearl, 1988] Judea Pearl. Probabilistic reasoning in expert systems: Networks of plausible reasoning, 1988.

[Pelletier and Elio, 1997] Francis Jeffry Pelletier and Renée Elio. What should default reasoning be, by default? *Computational Intelligence*, 13(2):165–187, 1997.

[Pelletier and Elio, 2002] Francis Jeffry Pelletier and Pelletier Renée Elio. Logic and cognition: human performance in default reasoning. In *In the scope of logic, methodology, and philosophy of science, Vol. I*. Citeseer, 2002.

[Pelletier and Elio, 2005] Francis Jeffry Pelletier and Renée Elio. The case for psychologism in default and inheritance reasoning. *Synthese*, 146(1-2):7–35, 2005.

[Penn *et al.*, 2008] Derek C Penn, Keith J Holyoak, and Daniel J Povinelli. Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2):109–130, 2008.

[Pfeifer and Kleiter, 2005] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human nonmonotonic reasoning. *Synthese*, 146:93–109, 2005.

[Pfeifer and Kleiter, 2009] N. Pfeifer and G. D. Kleiter. Framing human inference by coherence based probability logic. *Journal of Applied Logic*, 7:206–217, 2009.

[Pfeifer and Tulkki, 2017] N. Pfeifer and L. Tulkki. Conditionals, counterfactuals, and rational reasoning: An experimental study on basic principles. *Minds and Machines*, 27(a):119–165, 2017.

[Piaget, 1953] Jean Piaget. *Logic and Psychology*. Manchester University Press, 1953.

[Pinker, 2016] Steven Pinker. *The Blank Slate*. New York, NY: Viking, 2016.

[Plous, 1993] Scott Plous. *The psychology of judgment and decision making*. McGraw-Hill, 1993.

[Polberg and Hunter, 2017] Sylwia Polberg and Anthony Hunter. Empirical methods for modelling persuadees in dialogical argumentation. In Ieee International, editor, *Proceedings of the 29th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'17)*, pages 382–389. IEEE Press, 2017.

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

[Polberg and Oren, 2014] Sylwia Polberg and Nir Oren. Revisiting support in abstract argumentation systems. In Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti, editors, *Proceedings of the 5th International Conference on Computational Models of Argumentation (COMMA'14)*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 369–376. IOS Press, 2014.

[Pollock, 2001] John L. Pollock. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133(1-2):233–282, dec 2001.

[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.

[Prakken, 2019] Henry Prakken. Modelling accrual of arguments in ASPIC+. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 103–112, 2019.

[Ragni *et al.*, 2016] Marco Ragni, Christian Eichhorn, and Gabriele Kern-Isberner. Simulating human inferences in the light of new information: A formal analysis. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, pages 2604–2610, 2016.

[Rahwan and Tohmé, 2010] Iyad Rahwan and Fernando Tohmé. Collective argument evaluation as judgement aggregation. In Wiebe van der Hoek, Gal A. Kaminka, Yves Lespérance, Michael Luck, and Sandip Sen, editors, *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pages 417–424. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[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.

[Reznitskaya *et al.*, 2007] Alina Reznitskaya, Richard C. Anderson, and Li-Jen Kuo. Teaching and learning argumentation. *The Elementary School Journal*, 107(5):449–472, 2007.

[Riveret and Governatori, 2016] R. Riveret and G. Governatori. On learning attacks in probabilistic abstract argumentation. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'16)*, pages 653–661, 2016.

[Rosenfeld and Kraus, 2016a] 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.

[Rosenfeld and Kraus, 2016b] Ariel Rosenfeld and Sarit Kraus. *Strategical Argumentative Agent for Human Persuasion*. ECAI, 2016.

[Schellens *et al.*, 2017] Peter Jan Schellens, Ester Šorm, Rian Timmers, and Hans Hoeken. Laypeople's evaluation of arguments: Are criteria for argument quality scheme-specific? *Argumentation*, 31(4):681–703, Dec 2017.

[Schneider *et al.*, 2013] Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. Arguments about deletion: how experience improves the acceptability of arguments in ad-hoc online task groups. In Amy Bruckman, Scott Counts, Cliff Lampe, and Loren G. Terveen, editors, *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW'13)*, pages 1069–1080. ACM, 2013.

[Shanteau, 1989] James Shanteau. Cognitive heuristics and biases in behavioral auditing: Review, comments and observations. *Accounting, Organizations and Society*, 14(1-2):165–177, 1989.

[Silver and Veness, 2010] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Proceedings of the 24th Annual Conference on Advances in Neural Information Processing Systems (NIPS'10)*, pages 2164–2172, 2010.

[Simon, 1957] Herbert A Simon. *Models of Man: Social and Rational*. Wiley, 1957.

[Song and Ferretti, 2013] Yi Song and Ralph P Ferretti. Teaching critical questions about argumentation through the revising process: Effects of strategy instruction on college students' argumentative essays. *Reading and Writing*, 26(1):67–90, 2013.

[Song *et al.*, 2017] Yi Song, Paul Deane, and Mary Fowles. Examining students' ability to critique arguments and exploring the implications for assessment and instruction. *ETS Research Report Series*, 2017(1):1–12, 2017.

[Stanovich, 2003] Keith E. Stanovich. The fundamental computational biases of human cognition: Heuristics that (sometimes) impair decision making and problem solving. In Janet E. Davidson and Robert J.Editors Sternberg, editors, *The Psychology of Problem Solving*, page 291–342. Cambridge University Press, 2003.

[Stenning and Van Lambalgen, 2012] Keith Stenning and Michiel Van Lambalgen. *Human reasoning and cognitive science*. MIT Press, 2012.

[Thimm, 2012] M. Thimm. A probabilistic semantics for abstract argumentation. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI'12)*, FAIA. IOS Press, 2012.

[Thomas *et al.*, 2017] Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. Adapting healthy eating messages to personality. In *Proceedings of the 12th International Conference on Persuasive Technology (PERSUASIVE'17)*, pages 119–132. Springer, 2017.

[Thomas *et al.*, 2019a] Rosemary J. Thomas, Judith Masthoff, and Nir Oren. Can i influence you? development of a scale to measure perceived persuasiveness and two studies showing the use of the scale. *Frontiers in Artificial Intelligence*, 2:24, 2019.

[Thomas *et al.*, 2019b] Rosemary J. Thomas, Judith Masthoff, and Nir Oren. Is argumessage effective? a critical evaluation of the persuasive message generation system. In Harri Oinas-Kukkonen, Khin Than Win, Evangelos Karapanos, Pasi Karppinen, and Eleni Kyza, editors, *Proceedings of the 14th International Conference on Persuasive Technology (PERSUASIVE'19)*, pages 87–99, Cham, 2019. Springer.

[Tomasello, 1999] Michael Tomasello. The human adaptation for culture. *Annual Review of Anthropology*, 28(1):509–529, 1999.

[Tory Higgins, 2012] E. Tory Higgins. Regulatory focus theory. In P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins, editors, *Handbook of Theories of Social Psychology: Volume 1*, chapter 23, pages 483 – 504. Sage Publications Ltd, 2012.

[Toulmin, 1958] Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[Tversky and Kahneman, 1981] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[van Eemeren *et al.*, 2014] Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, Francisca A. Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. *Handbook of Argumentation Theory*. Springer, 2014.

[Varela *et al.*, 2016] Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind: Cognitive science and human experience*. MIT press, 2016.

[Villata *et al.*, 2017] S. Villata, E. Cabrio, I. Jraidi, S. Benlamine, M. Chaouachi, C. Frasson, and F. Gandon. Emotions and personality traits in argumentation: An empirical evaluation. *Argument & Computation*, 8:61–87, 2017.

[Walton *et al.*, 2008] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, NY, 2008.
[Walton, 1992] Douglas Walton. Rules for plausible reasoning. *Informal Logic*, 14(1), 1992.
[Warriner *et al.*, 2013] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.
[Wason, 1968] Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.
[Wu and Caminada, 2010] Y. Wu and M. Caminada. A labelling-based justification status of arguments. *Studies in Logic*, 3(4):12–29, 2010.
[Yu *et al.*, 2018] Zhe Yu, Kang Xu, and Beishui Liao. Structured Argumentation: Restricted Rebut vs. Unrestricted Rebut. *Studies in Logic*, 11(3):3–17, 2018.

Federico Cerutti
Department of Information Engineering, University of Brescia
via Branze, 38, 25123 Brescia, Italy
Email: federico.cerutti at unibs.it

Marcos Cramer
Faculty of Computer Science, TU Dresden
Nöthnitzer Straße 46, 01187 Dresden, Germany
Email: marcos.cramer at tu-dresden.de

Mathieu Guillaume
Centre for Research in Cognition and Neurosciences, Université Libre de Bruxelles
Avenue F. D. Roosevelt 50, 1050 Brussels, Belgium
Email: maguilla at ulb.ac.be

Emmanuel Hadoux
Department of Computer Science, University College London
Gower Street, London, WC1E 6BT, United Kingdom
Email: e.hadoux at ucl.ac.uk

Anthony Hunter
Department of Computer Science, University College London
Gower Street, London, WC1E 6BT, United Kingdom
Email: anthony.hunter at ucl.ac.uk

Sylwia Polberg
School of Computer Science and Informatics, Cardiff University
Queen's Buildings, 5 The Parade, Cardiff, CF24 3AA, United Kingdom
Email: polbergs at cardiff.ac.uk