Argument & Computation 14 (2023) 247–273
DOI 10.3233/AAC-220008
IOS Press

Confronting value-based argumentation frameworks with people's assessment of argument strength

Gustavo A. Bodanza [a,*] and Esteban Freidin [b]
[a] *Departmento de Humanidades, Universidad Nacional del Sur and Instituto de Investigaciones Económicas y Sociales del Sur, UNS-CONICET, Argentina*
E-mail: gbodanza@iiess-conicet.gob.ar
[b] *Instituto de Investigaciones Económicas y Sociales del Sur, UNS-CONICET, Argentina*
E-mail: efreidin@yahoo.com

Abstract. We reported a series of experiments carried out to confront the underlying intuitions of value-based argumentation frameworks (VAFs) with the intuitions of ordinary people. Our goal was twofold. On the one hand, we intended to test VAF as a descriptive theory of human argument evaluations. On the other, we aimed to gain new insights from empirical data that could serve to improve VAF as a normative model. The experiments showed that people's acceptance of arguments deviates from VAF's semantics and is rather correlated with the importance given to the promoted values, independently of the perceptions of argument interactions through attacks and defeats. Furthermore, arguments were often perceived as promoting more than one value with different relative strengths. Individuals' analyses of scenarios were also affected by external factors such as biases and arguments not explicit in the framework. Finally, we confirmed that objective acceptance, that is, the acceptance of arguments under any order of the values, was not a frequent behavior. Instead, participants tended to accept only the arguments that promoted the values they subscribe.

Keywords: Argumentation frameworks, value-based argumentation frameworks, argumentation semantics, human argumentation, experimental psychology

1. Introduction

Settling a dispute by arguing from different value-based viewpoints can sometimes be effective and sometimes sterile. For example, in the field of justice, it is common to see that a case well founded in evidence is lost due to arguments that point to procedural errors. The appeal is usually effective because for the audience of the courts of justice, the rules of judicial procedure have a higher value than the evidence. On the side of sterile argumentation, there are often dead end debates on the legalization of abortion, with pro arguments from, for example, the value frameworks of women rights and collective health, and con arguments from the frameworks of religion and the rights of the unborn. In this case, the argumentation tends to be sterile due to disagreements on the superiority of one value framework over the other.

Value-based argumentation frameworks (VAFs) [11,13,14,17,19,20,33] are computational models of persuasion intended to serve particular areas of practical reasoning and decision-making such as ethics

*Corresponding author. E-mail: gbodanza@iiess-conicet.gob.ar.

and law. Drawing inspiration from Chaim-Perelman's rhetorical theory, Bench-Capon [13,14,17,19] considered the strength of arguments as varying according to the values they promote and the assessment of those values by the particular audience they address. An argument is defeated if it is attacked by another argument promoting a value which is at least as preferred (by the audience) to the value it promotes. However, the argument is strong enough to withstand attack if the value it promotes is preferred. Then, argument interactions through an attack relationship can be analyzed to find the best justified arguments for the particular audience. This intuition facilitates the prescription of normative guidelines for the computational treatment of persuasion through argumentation.

Nonetheless, we could also think of VAFs as a theory about how persuasion occurs in real people, and psychological experiments could provide empirical support for the underlying intuitions. The broad approach we are proposing is not new. Bench-Capon's model is based on Dung's more general abstract argumentation frameworks [31], about which some empirical studies have been conducted to test certain "principles" and semantics for argument acceptance as common intuitions of ordinary people (see Section 2). However, as far as we know, the approach for the special case of VAFs is novel.

In this paper we reported a series of experiments carried out to basically confront the underlying intuitions of VAFs with those of ordinary people. We emphasize that the objective is *not* to test VAFs as an Artificial Intelligence argumentation model for persuasion, since, from a normative point of view, human reasoning cannot, unfortunately, be considered a standard of correct reasoning (in general, logical theories propose criteria of correct –sound– reasoning by reference to formal semantics –paradigmatically, for deductive reasoning–, but not by reference to actual human reasoning). In any case, we can still ask ourselves if, adopting a descriptive point of view, the same model is capable of explaining and predicting behaviors related to the practical argumentative reasoning of real people. Furthermore, on the way to finding an answer, empirical data could suggest hypotheses to enrich the model, either by incorporating or modifying elements that allow more and better applications.

The experiments we described here enabled us to observe that people's argument acceptance is more correlated with preferences among the values that the arguments promote than with specific forms of interaction among the arguments through an attack relationship. Moreover, the results showed that the relative importance of the values is assessed with different degrees, and the same argument can promote various values with diverse strengths. This is in line with Bench-Capon's recent claim [16] that a single ordering on values does not adequately capture the argument strength if there are several potential audiences. We also found that the evaluation of the interaction among arguments can be modulated by framing effects, and people usually change their perception on the relative importance of values depending on biases that incline their feelings towards conclusions. This finding is consistent with well-known bias effects reported by Kahneman, Tversky and others in the Heuristics and Biases program (e.g., [35,37], etc.).

The article is structured as follows. In Section 2 we review related works on empirical approaches to argumentation frameworks. Section 3 summarizes the formal definitions of VAFs and their semantics. In Section 4, we describe four experiments, analyzing and discussing their results individually. Finally, a general discussion and concluding remarks are offered in Section 5.

2. Related work

The approach of testing computational argumentation systems as models of human argument assessment has not been much explored, but is gaining increasing interest. VAFs are build on Dung's argumentation frameworks [31], extending the model by adding a set of values and a function that assigns

one value to each argument. The first study on Dung's argumentation frameworks model as a descriptive theory was reported by Rahwan and colleagues [45]. The aim was to test the *reinstatement* principle, according to which a defeated argument is reinstated if all of its attackers are in turn defeated. The authors conducted experiments that showed that an argument that is acceptable when it has no defeaters is also accepted after it has been defeated and restored, but with a lesser strength. This result was in agreement with Dung's semantics in terms of acceptability, but at the same time suggested that the model would be more in line with human common sense if it incorporated different degrees of confidence.

Cerutti et al. [25] found that the acceptability of arguments in human subjects corresponds mostly to the skeptical semantics of Prakken and Sartor's model [44]. However, they also observed important deviations that seem to arise from the implicit knowledge of subjects about the domain in which the evaluations are carried out.

Rosenfeld and Kraus [46] identified problems with more basic intuitions, such as conflict-freeness. All Dung's extension semantics satisfy this property, which indicates that the set of arguments accepted by a rational agent has no internal conflicts, that is, that the accepted arguments do not attack each other. Nevertheless, the authors found a significant percentage of surveyed individuals who accepted conflicting arguments, which they explained in part by the fact that some people try to appear impartial and good referees, capable of considering the different positions. They also assumed that some subjects may not accept defeat by arguments that promote a value regarded as weaker than that of the attacked argument (in line with the Bench-Capon value-based model, which is a specific object of study in this paper). In a later work [47], the authors analyzed the argumentative behavior of more than 1000 individuals (on different discussion subjects such as death penalty, flu vaccination or jury trials) and, using machine learning algorithms, found that it is possible to predict the choice of arguments with a high percentage, especially by combining a relevance heuristic based on distance measures among arguments. They then contrasted those results with Cayrol and Lagasquie-Schiex's bipolar argumentation theory [24] (an extension of Dung's model in which, in addition to an attack relationship, another support relationship is included), identifying that its predictions are not adequate.

Polberg and Hunter [41] also noted defects in Dung's model in explaining human argumentative behavior, and considered the uncertainty arising from subjects' opinions about both the arguments and the structure of the framework's graph to be of crucial importance. In relation to more optimistic results, Toniolo, Norman and Oren [50] found good predictions in the probabilistic semantics model of Thimm [49] and Hunter and Thimm [36]. These semantics estimate the probability of accepting an argument. First, the probability that a set of arguments is an extension for a given semantics is computed, and then the probability of accepting an argument is calculated as the sum of the probabilities of all those possible extensions to which it belongs. The experimental result was that people tend to agree with the semantics in terms of the credibility they attach to the conclusions of the higher probability arguments.

Cramer and Guillaume [30] argued that studies detecting a correspondence between human behavior and the semantics of argumentation frameworks are limited to contexts of a few arguments, hence such results cannot be generalized. Using more complex scenarios, they showed that part of the participants chose cognitively simple strategies that coincide with Dung's grounded semantics when analyzing strong acceptance (that is, arguments belonging to all the extensions), while others adopted more cognitively demanding strategies that are well predicted by CF2 semantics [7].

Recently, Bezou-Vrakatseli [23] replicated the findings of [45] on argument reinstatement, but obtained results contrary to some explanatory hypotheses about people's behavior.

To the best of our knowledge, no research similar to those described above has been conducted with respect to VAFs. There are, instead, several works in favor of *using* VAFs for the rational reconstruc-

tion of domain specific argument-based reasoning. For instance, Bench-Capon collaborated with several researchers in that line. With Atkinson and McBurney [18], they used VAFs to model economic experiments on the dictator game and the ultimatum game. With Bex and Atkinson [21], they combined VAFs with action-based alternating transition systems to represent the deliberations of characters of fables and how they weighed their values and motives given their attitudes. In that vein, the model was also used to understand narratives with argumentation [22]. In contrast, in the present paper we attempted to test the very intuitions underlying the main definitions of VAFs. Chorley and Bench-Capon [27] tested the value-based argumentation system developed by Bench-Capon and Sartor [10] from an empirical perspective, as a theory predictor for the explanation of real-life case-based legal reasoning. From this point of view, that work is more in line with ours than the others mentioned in this paragraph.

3. Value-based argumentation frameworks (VAFs)

We begin by summarizing the basic definitions of the model we are going to test. The primitive concepts are *arguments* and a binary relation that represents *attack* among arguments. These concepts are abstract, in the sense that there are no assumptions about the nature of the elements they represent or restrictions on their composition.

Definition 1 ([31])**.** An *argumentation framework* is a pair $AF = \langle AR, \ attacks \rangle$, where AR is a set of arguments and $attacks \subseteq AR \times AR$ is an attack relation among arguments.

Arguments interact through the attack relation. To determine which arguments survive such an interaction, different "semantics" can characterize the justification or warrant. As a result, each semantics sanctions a class of sets of argument, the *extensions* of AF, which can be thought of as the possible outcomes of the entire argumentation process.

Definition 2 ([31])**.** Given an argumentation framework $AF = \langle AR, \ attacks \rangle$, an argument $A \in AR$ is called *acceptable* w.r.t. a subset S of arguments of AR, if for every argument B such that B attacks A, there exists some argument $C \in S$ such that C attacks B. A set of arguments S is called *admissible* if each $A \in S$ is acceptable w.r.t. S, and S is *conflict-free*, i.e., the attack relation does not hold for any pair of arguments belonging to S. A set of arguments $S \subseteq AR$ is a *preferred extension* if it is a maximal (w.r.t. $\subseteq$) admissible set of arguments of AF. The *grounded extension* is the least (w.r.t. $\subseteq$) set $S \in AR$ such that $S = \{A : A \text{ is acceptable w.r.t. } S\}$.

Accepting arguments belonging to *some* preferred extension is usually considered the behavior of a *credulous* reasoner, while choosing arguments belonging to *all* preferred extensions is usually taken as the characteristic behavior of a *skeptical* reasoner. The grounded extension represents a (possibly) more skeptical behavior, since it is always a subset of every preferred extension.

Bench-Capon's VAFs are extensions of argumentation frameworks that take into account values associated with arguments. To that aim, a set of values and a function mapping arguments with values of the set are added.

Definition 3. A *value-based argumentation framework*[1] is a triple $\langle \langle AR, attacks \rangle, V, val \rangle$ where $\langle AR, attacks \rangle$ is an argumentation framework, V is a non-empty set of values, and val is a function

[1] Here we follow the lines of [19].

that maps from elements of AR to elements of V. For every $A \in AR$ and $v \in V$, if $val(A) = v$, then we say that A *promotes* value v. An *audience* a for $VAF = \langle\langle AR, attacks\rangle, V, val\rangle$ is identified with a preference order (irreflexive, asymmetric and transitive) $\succ_a \subseteq V \times V$.[2]

Different audiences for the same VAF may make dissimilar assessments of the arguments, according to their particular preference orders of values. Particularly, an attack by an argument A on an argument B is successful for an audience if, and only if, the value promoted by B is not preferred to that of A for that audience. Moreover, Bench-Capon's intuition indicates that if the value promoted by B is preferred to that promoted by A, then the conflict between A and B disappears, which (*ceteris paribus*) leads to the joint acceptance of both arguments. The following definition formalizes the semantics derived from that intuition. Note that the notion of *acceptability* reiterates that from Definition 2 with 'defeat w.r.t. an audience a' instead of 'attack,' while *admissibility* and the subsequent concepts are readjusted by reference to that new notion.

Definition 4.

- For arguments $A, B \in AR$, A *defeats (or is a successful attack on)* B w.r.t. audience a iff $(A, B) \in attacks$ and it is not the case that $val(B) \succ_a val(A)$.
- An argument $A \in AR$ is *acceptable* w.r.t. a subset of arguments $S \subseteq AR$ for an audience a iff for every argument B that defeats A w.r.t. a there exists an argument C that defeats B w.r.t. a.
- A subset of arguments $S \subseteq AR$ is *conflict-free* w.r.t. audience a iff for every $(A, B) \in S \times S$, either $(A, B) \notin attacks$ or $val(B) \succ_a val(A)$.
- A subset of arguments $S \subseteq AR$ is *admissible* w.r.t. audience a iff S is conflict-free w.r.t. a and for every $A \in S$, A is acceptable w.r.t. S for audience a.
- A subset of arguments $S \subseteq AR$ is a *preferred extension* w.r.t. audience a iff S is a maximal (w.r.t. $\subseteq$) admissible set w.r.t. audience a.
- An argument $A \in AR$ is *subjectively acceptable* iff there exists an audience a such that A belongs to some preferred extension w.r.t. a.
- An argument $A \in AR$ is *objectively acceptable* iff for every audience a, A belongs to every preferred extension w.r.t. a.

VAFs retain the abstract character of Dung's argumentation frameworks, which facilitates the instantiation of several structured models of argument. For instance, the ASPIC+ framework [43] and the DeLP system [34] are expressive enough to specify the internal structure and content of arguments, and allow for the use of explicit preferences to resolve attacks. Moreover, VAFs can instantiate argument schemes for practical reasoning, which is useful in representing legal case-based reasoning and public policy-making [4].

3.1. Two variations on VAF

As we carried out the experiments reported below, some hypotheses arose about variations on VAFs that could account for the obtained results. For the sake of readability, we introduce those variations in this section and will refer to them in due course.

[2] From here on, we refer to an arbitrary but fixed audience $\succ_a$, unless otherwise specified.

3.1.1. MVAF

The first variant concerns binding each argument to multiple values. This leads us to define the following model:

Definition 5. A *multi-value-based argumentation framework* (MVAF) is a tuple $\langle\langle AR, attacks\rangle, V, vals\rangle$, where $\langle AR, attacks\rangle$ is an argumentation framework, V is a set of values, and $vals : AR \to 2^V$.

Unlike VAFs, this model allows representing arguments that promote several values. Accordingly, we modify the notion of defeat as follows:

Definition 6. Let $\langle\langle AR, attacks\rangle, V, vals\rangle$ be a MVAF. For all arguments A, B $\in AR$, A *defeats* B w.r.t. audience a iff (A, B) $\in attacks$ and there is no $v \in vals$(B) such that for all $v' \in vals$(A) $v \succ_a v'$.

In other words, A defeats B if A attacks B and no value promoted by B is preferred to all the values promoted by A. This means that if B promotes some value that is preferred to all the values promoted by A, then the attack does not succeed.

3.1.2. QMVAF

The second variant refers to the possibility of assigning diverse importance degrees to the values, so that each argument is assigned a strength that results from the sum of the importance degrees of all the values that the argument promotes.

Definition 7. A *quantitative multi-value-based argumentation framework* (QMVAF) is a tuple $\langle\langle AR, attacks\rangle, V, importance, strength\rangle$, where

(1) $\langle AR, attacks\rangle$ is an argumentation framework,
(2) V is a set of values,
(3) for each $v \in V$, $v : AR \to \mathbb{R}^+$,
(4) $importance : V \to \mathbb{R}^+$, and
(5) $strength : AR \to \mathbb{R}^+$, such that for all X $\in AR$, $strength(\mathrm{X}) = \sum_{v \in V} v(\mathrm{X}) \times importance(v)$.

In words, (3) means that v(X) is the degree to which X promotes the value v, (4) means that $importance(v)$ is the degree of importance conferred to value v by the audience, and (5) means that the strength of X is calculated as a function of the degree to which X promotes each value multiplied by the importance given to that value.

The notion of defeat based on the strength of arguments is defined as follows:

Definition 8. Let $\langle\langle AR, attacks\rangle, V, importance, strength\rangle$ be a QMVAF. For all arguments A, B $\in AR$, A *defeats* B iff (A, B) $\in attacks$ and not $strength(\mathrm{B}) > strength(\mathrm{A})$.

Simply put, A defeats B if A attacks B and B is not "stronger" than A.

4. Experiments

All the experiments reported here were conducted under a general protocol approved by the Ethics Committee of the City Hospital of Bahía Blanca, Argentina. We recruited students from several groups for each experiment and no participant had a background in logic or argumentation theory.

We were mainly interested in finding out if people's assessments of arguments in a value-based argumentation scenario are related to their perceptions about attack among arguments, defeat and value preferences, consistently with Bench-Capon's intuitions. The first experiment was exploratory, and its results enabled us to propose a set of hypotheses that we tested in subsequent experiments.

4.1. Experiment #1

Participants were 64 voluntary undergraduates from different fields of study at the Universidad Nacional del Sur, Bahía Blanca, Argentina. Since we had no hypotheses regarding either gender or age, we did not collect data on those variables in these experiments. However, we knew it was a balanced representation by gender, while the vast majority of participants were between eighteen and twenty years old.

The experiment was conducted by means of an online single-variant Google form. We used a scenario similar to the one examined in [14],[3] with the following formulation:[4]

> *Martina, a diabetic, loses her insulin in an accident outside her responsibility. Before falling into a coma, she runs to Carla's house, an acquaintance who is also diabetic, to ask her for insulin. However, Carla is not at her house. Desperate, Martina decides to go in and use Carla's insulin.*
>
> *In a debate on this situation the following arguments are presented:*
>
> *A: Martina can go into Carla's house and use Carla's insulin, because her life depends on it.*
>
> *B: Martina can't do that, because that infringes on Carla's property.*
>
> *C: Martina can do that if she replaces Carla's insulin after the emergency.*
>
> *D: Martina can enter Carla's house and she does not have to replace Carla's insulin, because she could take it to save her life even though she is too poor to compensate.*

First, we asked to identify the values promoted by each argument by choosing an option among "Right to life," "Right to property," "Right to life and right to property," and "Neither." Second, we requested to express preferences between the values regarding the situation in question, by comparing them in terms of "more important than," "equally important," and "I don't know."

Then, participants were asked to identify compatibilities and incompatibilities between the arguments, instead of detecting attacks. This is because we assumed that, for an audience not specialized in argument systems, the term 'attack'[5] could have intentional or pragmatical connotations that are difficult to perceive, beyond the fact that one argument can be used to oppose another only if it is somehow incompatible with it. In that sense, we considered that incompatibility is a necessary condition for attack, since the attacked and the attacker cannot stand together on a reasonable basis. Hence, we asked the question below:

> *Which arguments do you consider to be, in some sense, incompatible with each other, that is, for some reason they cannot be accepted together? Choose the options with which you agree*:

followed by all the pairs of different arguments, with the options *compatible* and *incompatible* in a drop down menu. Another question was included with the aim of capturing the participants' intuition about what defeat between arguments means:

[3] This scenario, introduced by Coleman in [28], was also discussed in [13] but with a different representation.

[4] The text was in Spanish. The characters in the original story are Carla and Hal, but we changed to two female characters to avoid the possible influence of a gender bias.

[5] In Spanish, *ataque*.

If we put the arguments against each other, without taking into account the other arguments, do you think that one, by itself, defeats the other? Check the options with which you agree:

followed by all the ordered pairs of different arguments, with a tick box in each one.

Finally, participants were requested to evaluate arguments and conclusions. For arguments, we asked the following question:

Taking into account only the arguments that were used (A, B, C and D), check the argument(s) that, in your opinion, win(s) in this debate.

This question was intended to confront the accept/reject dichotomy of extension semantics. Then we aimed to check the correspondence among argument acceptance/rejection and the degrees of acceptance/rejection of relevant conclusions:"Martina can enter Carla's house and take her insulin," "Martina must not enter Carla's house and take her insulin," "Martina must replace Carla's insulin," "Martina does not have to replace Carla's insulin." In each case, we used a five-point Likert scale, ranging from "1 = Totally disagree" to "5 = Totally agree".

Results. The scenario was presented by Bench-Capon in [14] as the VAF depicted in Fig. 1. He identifies A and D with the value of life, and B and C with the value of property. Our results were analyzed in comparison with that presentation. All percentages are rounded.

We start by checking the perception of defeat. Figure 2 shows the percentages of participants that perceived each defeat and, among them, those who considered that both conditions were satisfied. Present participants' intuitions about defeat were fairly consistent with Bench-Capon's notion in some cases, but not in others. By definition, defeat implies attack (at least, incompatibility, following our hypothesis) and non-preference of the value promoted by the defeated argument, but this implication was only verified to a good extent for the pairs of arguments (A, B), (B, A), (C, B), and (C, D). Overall, 238 defeats were

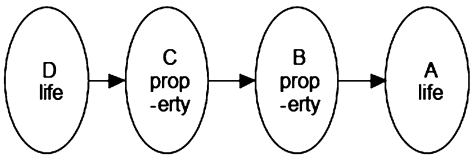

Fig. 1. VAF of the insulin scenario according to Bench-Capon (taken from [14]).

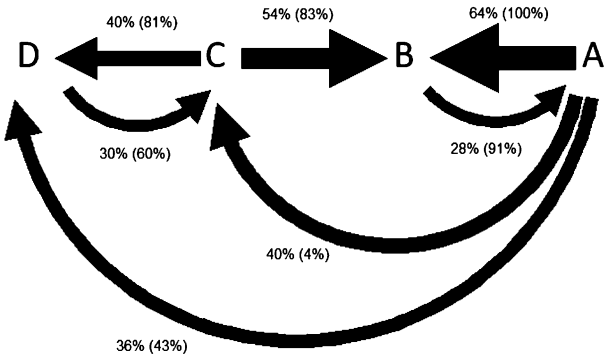

Fig. 2. Percentage of individuals who perceived each defeat. In parenthesis, percentage of participants who also observed both incompatibility between the arguments and non-preference for the value promoted by the defeated argument. For simplicity, we recorded only those defeats (arrows) observed by at least 20% of individuals.

Table 1
Experiment #1. Perception of values promoted and argument acceptance

	Right to life only	Right to property only	Right to life *and* right to property	Acceptance
A	**72%**	0%	17%	59%
B	1.6%	**94%**	1.6%	25%
C	**38%**	16%	31%	67%
D	11%	1.6%	**12%**	1%

reported, and participants identified an incompatibility between the arguments in tension in 159 of those 238 cases (67%). Therefore, most participants' opinions are consistent with the notion that people should perceive incompatibility between a defeated argument and its attacker. Having said that, it is noteworthy that a significant minority of judgments (33%) did not conform to that premise. In addition, it is clear that the representations of the scenario as a VAF that arise from the opinions of participants are mostly at odds with the Bench-Capon's representation given in Fig. 1.

As reported in Table 1, the experiment allowed us to observe that people quite clearly identify the values in agreement with Bench-Capon's opinion when they are explicit in the premises (arguments A and B), but not when they are not (arguments C and D). Of the total number of participants, 72% linked argument A exclusively to the right to life, and 94 % related argument B exclusively to the right to property. The case of argument C ("Martina can do that if she replaces Carla's insulin after the emergency") is surprising. In our opinion (and in line with Bench-Capon), it promotes the value of private property, but only one participant associated the argument exclusively with that value, while 31% identified it with both life and property rights, 30% did not link it to any of those values and, strangely, 38% related it *only* to the value of life. Maybe the answers of the last two groups can be explained by the fact that C *defends* or *reinstates* argument A, which is clearly associated by most participants with the value of life. For its part, argument D was identified with the value of life only by 11% of the participants, and with both values by 12%, while just one subject related it to the value of property only. This shows some tendency to associate the argument with the value of life, as expected, but the similar connection with both values at the same time is striking. Moreover, most of the participants (73%) did not respond, which can be interpreted as a difficulty in identifying the value promoted by D.

Opinions on winning arguments are summarized in Table 2. The "extension" {A, C} was the most accepted choice ($n = 20$, 31%). Within this group, the majority considered –curiously– that A promotes the values of *both* life and property. Argument C alone was the second most accepted option ($n = 15$, 23%), and the majority within this group determined, as expected, that it promotes the value of life. The third most accepted choice was A alone ($n = 7$, 11%), and here again the majority believed that it promotes the expected value, life. From here down the table, as we find smaller minorities, there are also more seemingly irrational answers. B was selected by 9% of the participants, the majority of whom judged that it promotes the value of life, contrary to expectations. However, nobody within that group stated that life is more important than property. The remaining choices were made by too few participants and with even rarer answers to draw any conclusions (for this reason, we did not show the data in the table). All in all, we can see some correlations between the arguments judged to be winners with the value they promote and value preferences, suggesting that the latter is a good predictor of argument choice, irrespective of other variables.

Regarding the conclusions, the responses were consistent with the selection of arguments. The acceptance (resp., non-acceptance) of an argument given the answers "agree" or "totally agree" (resp., "totally disagree," "disagree," or "neither agree nor disagree") with respect to its conclusion was observed in 86%

Table 2
Experiment #1. Arguments chosen as winners and values promoted according to majority judgment, and claimed value preference

Winner/s ("extension")	Acceptance	Value promoted acc. majority opinion ((L)ife, (P)roperty, (U)nknown)	Value preference		
			Life	Equal	Property
A, C	31%	L and P, L	60%	40%	0%
C	23%	L or U	22%	78%	0%
A	11%	L	100%	0%	0%
B	9%	P	0%	67%	33%
A, B	7%	L or U, P	0%	60%	40%
B, C	5%	P, L or U	0%	100%	0%

of the cases (221 out of 256). Conversely, the agreement (resp., disagreement) with a conclusion given the acceptance (resp., non-acceptance) of its supporting argument was observed in 98% of the cases (251 out of 256). The difference seems to make sense given that, in some situations, people are inclined to accept a conclusion even if they do not accept some supporting arguments.

Next, we assessed the extent to which VAF can be used to make correct predictions about argument acceptance and compared that with other variables. In each case, we took into account perceived defeats, perception of what values the arguments promote, and value preferences. We checked the following variants:

- VAF (Definitions 3 and 4). We verified only those cases in which each argument was associated with *at most* one value. Since only three participants in this experiment indicated that each argument promoted *exactly* one value, we decided to also count those cases in which no value was indicated for some arguments. The arguments that were not associated with any value were treated as if they promoted some value indifferent to the other values. For example, assume that a participant indicated that X attacks Y, but some of those arguments were not identified as promoting any value. Whatever the case, we counted a defeat from X to Y, since the argument that was not associated with any value was treated as promoting an indifferent value with respect to the one promoted by the other argument.
- VAF (I). It could be the case that the participants' intuitions about 'defeat' did not coincide with the expected interpretation. If this is the case, the appreciation of incompatibility (I) between the arguments in that relationship would suggest a more precise correspondence. Consequently, we checked if considering defeat together with incompatibility perceptions made a significant prediction difference with respect to considering defeat perceptions alone.
- MVAF (Definitions 5 and 6) In this model, arguments can be perceived as promoting *more than one* value. Just like we did with VAF, we also counted those cases in which some arguments were not recognized as promoting some value.
- MVAF (I). The same as MVAF, but taking incompatibility as a reinforcement of defeat.
- AF (Definitions 1 and 2). Dung's simple argumentation frameworks model. We intended to know to what extent that model predicts differently than VAF and MVAF. Here, we assimilated defeats to attacks, regardless of value assessments.
- AF (I). The same as AF, but taking incompatibility as a reinforcement of defeat perceptions.
- V. Finally, we verified to what extent the promoted value and the preference for that value, regardless of any other variables, is enough to make correct predictions.

Table 3
Experiment #1. Percentages (mean) of successful predictions with confidence intervals

Model	Percentage (95% CI)
VAF	69.6 (65.3, 73.9)
VAF (I)	68.8 (62.8, 74.7)
MVAF	66.8 (63.6, 70.0)
MVAF (I)	68.7 (65.9, 71.6)
AF	66.4 (63.2, 69.6)
AF (I)	68.7 (65.8, 71.7)
V	71.88 (69.1, 74.6)

Results are displayed in Table 3, which shows that differences are in a small range of around 5.5 percentage points. As can be seen, VAF worked relatively well, but the best performance was obtained by taking into account only the promoted values and value preferences. Beyond mean differences, the overlap of the confidence intervals evidences the closeness of the predictive efficacy among all models.

Discussion. We observed that the association of arguments and values was very heterogeneous with the exception of argument B, which almost all participants perceived as promoting the value of property. The case of argument C, which 31% of the participants linked to both values, evidences the need to consider that arguments can promote more than one value in order to, at least, get more accurate representations. In sum, the association of arguments with values seems to be influenced by subjective factors, which poses a representational issue, and different values can converge in one single argument, unlike VAFs' underlying intuition. Bench-Capon himself was aware of these limitations and identified certain important factors to take into account in some specific domains, such as the authorities the audience respects, their attitude to risk and their degree of loss aversion [16]. In fact, it is also interesting to investigate a variety of biases as factors that modulate value preferences.[6]

A different source of doubt regarding the association of arguments with values lies in a possible inadequate design of our experiment. In view of the results, the option "right to life *and* right to property" to identify the promoted values could have induced some guessing responses in case of doubt or confusion. This was taken into account for the next experiment.

We compared VAF-based predictions with several variants, obtaining similar results. However, the best performance was that of the model that considers value preferences as the only relevant variable. This suggested that improved measurement of appreciations relative to values could give us better differentiated results. Likewise, we suppose that intuitions about defeat might be better tested in simpler scenarios where attacks are more clearly one-way directed and value promotion is more easily identifiable.

We can also analyze other sources of representational problems. The directionality of the attack relation is certainly a very important issue. Bench-Capon himself was aware of the problem and he did not have a definite solution. In fact, the insuline scenario, which he represented in [14] by a linear relation where D attacks C attacks B attacks A (the representation we used in our study), was instead represented in [13] by a 3-cycle where A attacks C attacks B attacks A. Moreover, in [12], the author said the following:

> Consider the trite example of the arguments 'Kerry can fly because she is a bird' and 'Kerry cannot fly because she is a kiwi.' Here, although we have contradictory conclusions, we would naturally say that the second

[6] A reviewer also suggested to investigate logical fallacies that could affect the perception about values.

> argument attacks the first, but not vice versa. Sometimes, therefore I shall represent what might in logic be considered a mutual attack as an attack in one direction only, when it seems clear that this is how the arguments are intended. By giving ourselves this amount of freedom, we should be able to construct the most natural representation possible.

In sum, though the directionality of attacks is hard to define, how arguments are intended could be considered to decide the right representation (we will discuss more on the directionality of attacks in Section 5). In Experiments #3 and #4 below, we used natural language arguments that more clearly suggested a one-way attack intent.

Another sign of a possible representation issue is the low acceptance of argument D (and of its conclusion). This might suggest the intervention of an unspoken argument in the minds of the participants, one that defeats D. Given that D considers that a person with limited economic resources would have the right to take the insulin without the obligation to compensate the owner, the argument could be interpreted as referring to exceptional cases, and that the information in the context does not indicate that Martina's case fits such an exception. This suggests, at least, three representational rectifications of the models. One is to take into account the incidence of some argument E attacking D, based on considerations of irrelevance for the case in question. This fifth argument would promote, in turn, the value of information, or the lack of it, prevailing over the others. The argument D can raise some critical questions that can be expressed in the form of arguments, and that can be represented through structured argumentation versions of VAF (see, for example, [2,3,18], and [8] for representing exceptions). A second rectification consists of simply deleting D from the representation, while a third one consists of considering an attack from C to D. All these rectifications would explain the high esteem for argument C, whose only attacker, D, would be ignored or rejected.[7]

These possibilities are in line with [26], where the authors claimed that, in order to create argumentation systems, designers must take into account implicit domain-specific knowledge or beliefs. Maybe these representational problems are more important than the extent to which the empirical data match some semantics,[8] in the sense that they call for a foregoing resolution. In the meantime, these problems could be tackled with experiments and questionnaires specifically aimed at elucidating what is being represented, and/or using simpler scenarios whose representations raise fewer doubts.

In view of the above results, we planned to test the following hypotheses:

- H1: The strength of arguments is assessed as a function of the various degrees of importance conferred to the values the arguments are perceived to promote (not just as a function of a preference order), and different values can converge in the same argument with dissimilar strength degrees.
- H2: When deciding on a given debate, people tend to consider unspoken arguments.
- H3: The preference of the value promoted by an attacked argument may have the effect of avoiding defeat, but the attacking argument is not acceptable together with the attacked one since the conflict persists.
- H4: Acceptance of value-based arguments can be modulated by biases and framing effects.

Hypotheses H1 and H2 were tested in Experiment #2, while H3 and H4, in Experiment #3 and Experiment #4, respectively.

[7] Note that those who considered that A attacks C tended not to see any incompatibility between these arguments (Fig. 2).
[8] This opinion was subscribed by an anonymous reviewer.

Table 4
Experiment #2. Perception of values promoted and argument acceptance

	Right to life only	Right to property only	Right to life *and* right to property	Acceptance
A	**92%**	0%	0%	52%
B	1.6%	**94%**	0%	13%
C	**44%**	21%	27%	59%
D	**69%**	8%	3.2%	4%

4.2. Experiment #2

Sixty-two undergraduate students from different fields of study were voluntarily recruited at the Universidad Nacional del Sur, Bahía Blanca, Argentina. In order to test hypotheses H1 and H2, we maintained the questionnaire from the first experiment, except for the questions that involved the variables relevant to the hypotheses, namely, association of arguments with values, varying degrees of importance assigned to values, and possible influences of external arguments. Regarding hypothesis H1, participants were asked to express their beliefs about the value(s) each argument promotes. To answer, they had to mark cells in a table with four rows headed by the names of the arguments (A, B, C, D) and two columns titled "right to life" and "right to property". Then, participants were requested to indicate the importance of the values with respect to the referred situation (the insulin case), each one on a 10-point scale where 1="Not important at all," and 10="Absolutely important." Regarding hypothesis H2, participants were asked to answer the following question:

> *If you think there is an important argument missing from this debate, please write it down. Which of the arguments seen would that argument attack/support?*[9]

Results. Table 4 summarizes the results about the perception of the values promoted by the arguments. Our change of strategy in presenting of answer options seemed to make a significant difference in the assessment of argument D, which in this experiment was associated with only the value of life by 69% of the participants. We also found that A and C were more clearly linked to the value of life only (in the case of A, no participant identified it with both values).

To test H1, we formulated the hypothesis in a more precise way by taking into account the following operationalization of functions, where X $\in$ {A, B, C, D} and i is a participant:

- $life_i(X) := 1$, if i perceived X as promoting the value of life; 0, otherwise;
- $property_i(X) := 1$, if X is perceived as promoting the value of property; 0, otherwise;
- $importance_i(value) := g$, where $g \in \{1, \ldots, 10\}$, and $value \in \{life, property\}$ (i.e., the importance given by i to *value* in a ten-point scale);
- $strength_i(X) := life_i(X) \times importance_i(life) + property_i(X) \times importance_i(property)$ (i.e., the strength of X for i is given by the sum of all the importance degrees assigned to the values promoted by X according to i's opinion)

Then, we reformulated the hypothesis as follows:

- H1: The set of arguments accepted as winners by the individual i is

$$W_i \stackrel{\text{def}}{=} \{\text{X} : \text{for all Y} \in \{\text{A, B, C, D}\},\ strength_i(\text{X}) \geqslant strength_i(\text{Y})\},$$

[9] Unlike bipolar argumentation models [24], supporting arguments do not play a role in VAFs. However, if the participants expressed supporting arguments, then we would have, on the one hand, a better understanding of their behavior and, on the other hand, some evidence that would show the limits of VAFs to correctly represent the situation.

Table 5
Experiment #2. Percentages (mean) of successful predictions with confidence intervals

Model	Percentage (95% CI)
VAF	69.4 (66.2, 72.7)
VAF (I)	55.6 (49.2, 61.9)
MVAF	67.7 (65.0, 70.5)
MVAF (I)	56.0 (53.5, 58.6)
AF	68.5 (66.0, 71.1)
AF (I)	56.4 (54.0, 59.0)
V	62.9 (60.2, 65.6)
QV	65.3 (62.1, 68.6)
QMVAF	67.7 (64.9, 70.5)
QMVAF (I)	56.5 (53.9, 59.0)

i.e., the set of all the arguments whose strength is maximal w.r.t. $\geqslant$, according to i's criterion.

Then we analyzed a quantitative version QV of model V, that predicts the choice set W_i for each participant i. Like V, this model does not take into account attacks or defeats.

Additionally, we aimed to analyze the quantitative version QMVAF (Definitions 7 and 8), with the variants of considering defeat perceptions, on the one hand, and defeat reinforced with incompatibility perceptions (QMVAF (I)), on the other.

The comparison of correct prediction rates is displayed in Table 5. As in Experiment #1, we tested VAF-based predictions only in those cases in which the arguments were perceived as promoting at most one value ($n = 45$).

Regarding H2, i.e., the hypothesis that participants considered the incidence of external arguments, 19 participants (31%) introduced new arguments or, at least, some comments, either for explaining their decisions or for criticizing some point of view. Among them, we identified 11 out of 19 arguments either presenting objections to argument D or supporting the contrary conclusion, i.e., Martina must replace the insulin. Here are some examples:

> *Martina can verify that she has her phone with her to call Carla. That way she could ask his permission to take her insulin and then give it back.*
>
> *Martina could enter the house if, and only if, her life is in danger. The insulin would be stolen, she would have to replace it and abide by the force of the law.*
>
> *I would attack argument D, because if Carla also ran out of insulin, her life would also be in danger. Therefore, I consider that if Martina could NOT replace her insulin because she is poor, she would be consciously attempting the life of another person. She should, in any case, look for the means to replace that insulin regardless of her economic reach.*
>
> *Even if Martina was too poor, it should be considered how important it is to replace Carla's insulin, since she took away Carla's supply and it may also be essential for her to recover it again.*
>
> *I believe that, in support of argument C, Martina should replace the insulin quickly, since a similar situation can happen to her friend Carla even at the same moment, and the latter goes home quickly*

hoping to save herself, since she is convinced that she has insulin at home. And when she arrives she finds that it is not like that.

I think the situation is really much more complex than the one presented. Even the potential expressed in the conditional "if she were too poor" appears as contradicting that she lost her insulin at her house (somehow she had the resources to have it). There is also the question of permission and obligation. For me they are two different things. And the third point is that apparently they know each other (she knows that Carla has insulin) and there are other issues at stake, in addition to the fact that Carla's life would also be at stake.

In either case, we conclude that the participants who introduced those arguments considered some explicit or implicit defeat to D.

Discussion. We can see that VAF predicted comparatively quite well in this experiment. Now, two results are striking:

(1) When defeat and incompatibility were considered together to predict accepted arguments (I), predictions were less accurate than when defeat alone was taken into account (in contrast to results in Experiment #1). We could think that this issue is due to the fact that, when we considered defeat and incompatibility together to predict acceptance, we ended up establishing a stronger criterion for making a prediction (to wit, a perceived attack was dismissed if it was not accompanied by a perceived incompatibility between the attacker and the defeated argument). This criterion seemingly leads the models to avoid taking into account weak signals that have predictive power.

(2) Model V achieved a lower performance than in Experiment #1 (though it improved somewhat in the quantitative version QV). Unfortunately, our possible explanation for (1) does not explain (2). Alternatively, we could relate both issues to how we inferred the participants' preference among the values, which is the only methodological variation between the experiments. Whereas in Experiment #1, we directly asked them to choose whether they prefer one value over the other or were indifferent (i.e., a comparative framework), in Experiment #2 we requested the participants to numerically express the level of importance of each value separately (non-comparative framework). This might have weakened the predictive power of the values in Experiment #2, since that information affects validations of both defeat and value preference. Although we do not have an explanation of how the information difference could generate the effects of (1), it provides a good reason to account for (2). Indeed, some classic findings in experimental psychology (e.g. [38]) show that when individuals evaluate alternatives numerically, they do not necessarily make comparative judgments, and preferences inferred from those judgments may be reversed compared to more direct choices.

Arguments were more clearly associated with values than in the first experiment. Though, in some cases, a single value assignment tendency was appreciated, in concordance with Bench-Capon's model, some arguments were still associated to both values. The perception of the values promoted by arguments and the measure of their importance were still correlated with argument acceptance regardless other variables intrinsic to the model (such as attacks, defeats, conflict-freeness, or any usual extension semantics), which makes V and QV simple estimation methods for argument acceptance. Predictions include those cases that cannot be represented as VAFs because arguments are perceived as promoting more than one value. In cases that can be represented as VAFs and values can be ordered asymmetrically, Bench-Capon's model performs something better (67% vs. 58%). Still, V and QV are more parsimonious, in the sense that they suggest less complex computations with fewer variables.

4.3. Experiment #3

In the previous experiments, we were not able to determine on a good basis how participants perceived attack directions, independently of incompatibilities. That was in part due to the fact that the arguments' conclusions in that scenario were contradicting, which could suggest an attack in either direction. As Bench-Capon [15] noted, knowing the type of attack can be crucial.[10] In this experiment, we aimed to test the hypothesis that the preference of the value promoted by an attacked argument may have the effect of avoiding defeat, but the attacking argument is not acceptable together with the attacked argument since the conflict persists (H3).

Fifty-six volunteer students from different disciplines at the Universidad Nacional del Sur, Bahía Blanca, Argentina, participated in the experiment. Paper questionnaires were delivered by hand. Unlike the previous experiments, in the scenario of the present study we used different arguments A and B in such a way that B denies that the evidence can be used to condemn X. Therefore, B is an undercutting defeater of A. Moreover, in the wording we used, B begins with 'However', which suggests that B is a counterargument of A, hence, favoring a one-way interpretation of the attack:

A: X should be convicted, because the evidence shows that he was responsible for the crime.

B: However, the evidence was obtained illegally, so it cannot be used to convict X.

Note that we could not have gotten the desired interpretation if we had asked about an incompatibility instead of an attack. If participants were able to understand that B attacks A and not the other way around, then we would have a good basis for analyzing defeat as dependent on a correct perception of the values promoted and preference over those values, in accordance with VAF's underlying intuition. In addition, we asked about the comparative preference between the promoted values (as in Experiment #1, to avoid the possible effect that occurred in Experiment #2) and about any other possible values that the participants would consider to bias their choice. The questionnaire was as follows:

(1) *With no other information available, which of these arguments would you accept as the winner? (Options: Only A; Only B; Both; None; I would accept one, but I don't know which one)*
(2) *Do you consider that argument B is used to attack argument A? (Options: Yes; No; I don't know)*
(3) *Do you consider that argument A is used to attack argument B? (Options: Yes; No; I don't know)*
(4) *Do you agree that argument A promotes evidence as the main value? (Options: Yes; No; I don't know)*
(5) *Do you agree that argument B promotes the legality of the process as its main value? (Options: Yes; No; I don't know)*
(6) *Personally, what importance do you give to those values in this context (Options: The legality of the process is more important than the weight of the evidence; The weight of the evidence is more important than the legality of the process; Both are equally important; I don't know)*
(7) *Do you consider that there is any other value involved in this case that inclines your decision towards one argument or another? (Options: There is no other value; Yes, the value is . . . [participant should complete])*
(8) *(If your answer to question (7) was 'Yes') Which of the arguments do you consider promotes this value? (Options: A; B; Both; None)*

[10] An attack against a conclusion is known as a *rebutting* attack, while an attack against a premise is known as an *undermining* attack. A third kind is *undercutting*, where attack is directed against the connection between the premises and the conclusion [43]. Rebutting attacks give rise to symmetric attacks, while undermining and undercutting attacks do not (for structured arguments implementing these kinds of attacks in the context of value-based argumentation see [8].)

Table 6
Experiment #3. Tendencies in participants' opinions

Winner/s			B attacks A	A attacks B	A promotes evidence	B promotes legality	Value importance		
A	B	Both					Evidence	Legality	Equal
45%	27%	14%	**70%**	18%	**84%**	**71%**	30%	21%	48%

(9) *(If your answer to question (7) was 'Yes') What comparative importance do you give to that value in this context...*

a. *regarding the weight of the evidence? (Options: more important; less important; equally important; I don't know)*
b. *regarding the legality of the process? (Options: more important; less important; equally important; I don't know)*

Results. Results in Table 6 show clear tendencies to agree with the intuitions that the only attack is from B to A, that A promotes the value of evidence, and that B promotes the value of legality. This is in accordance with VAF's canonical representations. The other variables rely on more subjective appreciations.

Now, according to VAF, we have the following predictions. If participants only perceived the attack from B to A, then:

(i) if the value promoted by A (evidence) is not more important than the value promoted by B (legality), then B defeats A; hence, participants should accept B and reject A;

(ii) if the value promoted by A (evidence) is more important than the value promoted by B (legality), then B does not defeat A; hence, participants should accept *both* B and A.

In either case, B should be accepted, i.e., B should be *objectively* accepted. Moreover, if evidence and legality are perceived with the same importance, then VAF reduces to a Dung's AF, where the attacking argument is the winner. In turn, H3 predicts the same as VAF in (i) and the acceptance only of A in (ii).

In Table 7, we compared the VAF-based predictions and H3 with respect to (i) under two different conditions: (a) assuming the asymmetric attack from B to A as a fact, no matter what participants said about that, and (b) taking into account only cases in which participants recognized that asymmetric attack. In our data, the group that is crucially relevant to compare the hypotheses was formed by 27 participants who accepted that (1) A promotes evidence, (2) B promotes legality, and (3) evidence is preferred or indifferent to legality. According to VAF, the participants accepting these three conditions would still accept B, while H3 predicts the opposite (note that, if condition (3) is not fulfilled, then the acceptance of B would not contradict any hypothesis).

For (a), V predicts that individuals will not accept both A and B, which could be interpreted to mean that the arguments are anyway considered in conflict, in concordance with H3. Consequently, they should choose *at most* one argument. Results show that 23 (85%) participants accepted either only one argument or none: among the participants who preferred evidence over legality, 89% chose only A (8 out of 9), and among those who declared indifference between the values, 64% opted for only A, and 36% for only B (9 and 5 out 14, respectively). In contrast, only 6 individuals (22%) answered in accordance with the VAF-based prediction: 5 chose B when they were indifferent between the values, and 1 chose both A and B preferring evidence to legality.

For (b), H3 had a prediction efficacy of 81%: 13 out of 16 participants chose only one argument. Moreover, 5 out of 6 (83%) chose A when preferred evidence, and 57% (4 out of 7) and 43% (3 out of 7) chose A and B, respectively, when declared indifference between the values. In contrast, we registered a

Table 7
Experiment #3. Successful argument acceptance prediction given B′s attack on A

$n = 27$		VAF	H3
B attacks A	(a)	22%	85%
	(b)	25%	81%

Table 8
Experiment #3. Percentages (mean) of successful predictions with confidence intervals

Model	Percentage (95% CI)
VAF	49.1 (44.2, 54.0)
AF	44.6 (39.6, 49.7)
V	59.8 (54.9, 64.7)

VAF-based prediction efficacy in 25% (4 out of 16): 1 out of 7 (14%) participants opted for both A and B when they preferred evidence over legality, and 3 out of 9 (33%) participants chose B when declared indifference between the values. Finally, when legality was perceived as more important than evidence, and provided that participants identified A with evidence and B with legality, both VAF and H3 had identical (low) efficacy: both succeed in the same 4 samples out of 8 (50%) under condition (a), and in the same 2 samples out of 6 (33%) under condition (b). However, if we ignored how participants identified the arguments and the values and just observed the value-argument correlation assuming that A promotes evidence and B promotes legality, both VAF and H3 were 73% successful in predicting the choice of B when legality is preferred.

Next, we considered the acceptance predictions in the frameworks elicited by the participants' opinions, comparing VAF, AF, and V (Table 8). As can be seen, V is the model with the highest predictive accuracy and is the only one of the three that performs above chance level.

Regarding questions (7)–(9), we only obtained three answers, which is insufficient to shed any light on the behaviors.

Discussion. The results showed low agreement with the concept of objectivity proposed by Bench-Capon and replicated the participants' tendency to choose the argument that promotes the preferred value. Indeed, argument B, which VAF deems as objectively acceptable, only obtained 27% acceptance (regardless of other variables), much less than A, which was accepted by 45% of the participants (Table 6). In contrast, H3 predicted with much more success. This suggests that VAF could be adjusted by giving more importance to value promotion in order to better fit as a theory of human argument acceptance.

It can be reasonably argued[11] that the wording "is used to," when we asked about attacks, is not related to the conceptual assumptions of the model. In argumentation theory, an attack from X to Y generally means that X can be used as a counterargument against Y, not that X is actually used as a counterargument against Y in a given debate. Indeed, "is used to" may give rise to pragmatic considerations that are not at stake. Now, on the one hand, we believe that in the context of our experiment, either wording leads to a similar interpretation. Even if we used the wording "is used to attack", we think that the most straightforward interpretation of that phrase is that "it could be used to attack," because there were no subjects actually implementing the arguments to attack each other. On the other hand, suppose that the

[11] A reviewer raised this criticism.

wording suggests a biased interpretation of directionality that does not correspond to an adequate theoretical concept. However, 'attack' is usually defined in such a way that the conditions for it to occur are not subject to individual interpretations but to objectively verifiable logical/linguistic relationships. In accordance with the definition given in ASPIC+, for instance, it is the case that B attacks A, but A does not attack B, because B is an undercutting defeater of A. The notion of 'defeat' in VAF, in turn, depends on both an (objective or given) relationship of attack and a (subjective, dependent on the audience) preference over the promoted values. So, it was expected that B was accepted by most of the participants. In terms of VAF, B is *objectively* acceptable, because, in any case, B attacks A and either a) the value promoted by A is not preferred, or b) it is preferred, implying that B does not pose a threat to A. However, the experiment showed that –contrary to the VAF-based prediction– B was not objectively accepted by the participants. All in all, the small difference found between conditions (a) and (b) in Table 7 indicates that the perception of the direction of attacks is not determinant for argument acceptability.

Regarding the low prediction efficacy of both VAF and H3 when participants preferred legality to evidence, we can only provide speculative explanations. A possible psychological reason is that participants could declare preference for the "correct" value of legality, according to the rule of law, while their sincere feelings are either on the side of evidence or balanced between both values. More generally, the content of the arguments could induce the occurrence of biases and framing effects that modulate the expected incidence of value preferences. In particular, we used a scenario where conviction was under discussion, which could generate a leniency bias that affected the evaluation of the arguments. Experiment #4 is proposed to test this bias in different framings, as a classic challenge to normative effects on judgment and decision making [51].

Finally, the same outcome as that of model V could be attained by a VAF in which there is a bidirectional attack between A and B. So, a possible explanation for this outcome could be that the participants really feel the conflict to be a bidirectional attack, but do not indicate this in the questions about attacks, because they do not correctly understand the meaning of the specialist term 'attack' (or are misled by the words "is used to," as discussed before). As it can be seen, the problem of the directionality is recurrent.

4.4. Experiment #4

We recruited 74 undergraduate students from different disciplines at the Universidad Nacional del Sur and the Universidad Salesiana, Bahía Blanca, Argentina. The questionnaires were completed online in Google Forms. To test the hypothesis that the acceptance of value-based arguments can be modulated by biases and framing effects (H4), we designed two frames (scenarios) with identical representation as value-based argumentation frameworks and equal values, but inverting conclusions with opposites (acquit/convict):

Frame 1 (F1)**.** *A: X should be acquitted, because evidence E1 shows that he was not responsible for the crime.*

B: However, evidence E1 together with evidence E2 clearly shows that X was responsible for the crime.

C: But evidence E2 was obtained illegally, so it cannot be used to convict X.

Frame 2 (F2)**.** *A′: X should be convicted, because evidence E1 shows that he was responsible for the crime.*

B′: However, evidence E1 together with evidence E2 clearly shows that X was not responsible for the crime.

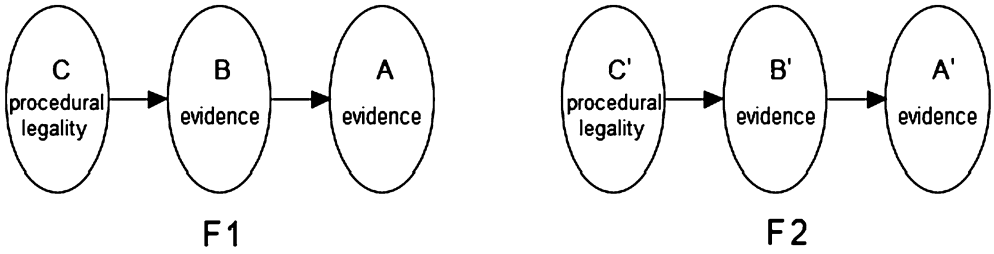

Fig. 3. VAFs with identical structure representing the frames used in Experiment #4.

C′: But evidence E2 was obtained illegally, so it cannot be used to acquit X.

The same questionnaire followed each frame:

(1) *With no other information available, which of these argument(s) would you accept as winning in this discussion? (Options: A; B; C; None; I would accept one, but I don't know which one)*
(2) *What arguments do you consider that promote evidence as the main value? (Options: A; B; C; None)*
(3) *What arguments do you think that promote the legality of the process as the main value? (Options: A; B; C; None)*
(4) *In your opinion, what importance do you give to these values in this context, on a five-point scale where 1 = minimum and 5 = maximum?*
Legality of the process: . . .
Evidence: . . .
(5) *Do you consider that there is some other value involved in this case that inclines your decision towards one argument or another? (Options: There is no other value; Yes, the value is . . .)*
(6) *(If your answer to question (5) was 'Yes') Which of the arguments do you think promotes this value? (Options: A; B; C; None)*
(7) *(If your answer to question (5) was 'Yes') What importance do you give to this value in this context, on a five-point scale where 1 = minimum and 5 = maximum?*

According to our intuition, A, B, A′, and B′ all promote the value of evidence, while both C and C′, the value of procedural legality. On the other hand, B and B′ pose asymmetrical (specificity-based) attacks against A and A′, respectively, while C and C′ generate asymmetrical (undermining) attacks against B and B′, respectively (Fig. 3).

We tested two conditions, C1 and C2, only varying the order of presentation of the frames: in C1 ($n = 35$), F1 was presented first and F2 second, while, in C2 ($n = 39$), the order was inverted.

One point to observe was to what extent responses about the second frame showed different intrasubject perceptions of arguments and/or values with respect to the first frame. One variable of interest was the "positional concordance" among the arguments chosen in one frame and the other, and we classified it into three classes: a) strict concordance: argument X was chosen in one frame iff X' was chosen in the other frame (i.e., the arguments chosen occupy the same position in their respective graphs); b) weak concordance: the selection of arguments changed regarding their positions but there is no conflict (attack) between their positions (i.e., from A to C′, or from C to A′); and c) non-concordance: the choice of arguments changed regarding their positions and there is conflict (attack) between their positions (i.e., from A to B′, from B to A′, from B to C′, or from C to B′).

Results. In both conditions C1 and C2, we observed similar percentages of concordance in general (differences are not statistically significant). We found 57% of strict concordance in C1 and 59% in C2, 17% of weak concordance in C1 and 15% in C2, 17% of non-concordance in C1 and 26% in C2 (in C1, there were 9% of participants who did not choose any argument in some of the frames, so we do not register that as a fact of concordance/no concordance). The greater tendency towards strict concordance would favor the robustness of the structural factors as predictors of choice under similar value preferences.

VAF predicts the following behaviors regarding acceptability. If the values are assessed as having the same importance, then the prediction matched that of Dung's model, that is, to chose A and C, and A′ and C′, for F1 and F2, respectively. If evidence is evaluated as being more important than procedural legality, then the prediction is choosing B and C, and B′ and C′, in F1 and F2, respectively. And if procedural legality is assessed as more important than evidence, then the prediction is to opt for A and C, and A′ and C′, in F1 and F2, respectively. The results show that the effectiveness of these predictions in the intrasubject analysis is zero, if we regard the exact extension as the only successful prediction. However, it is around 35% (37% in C1 and 33% in C2) if we consider choosing *some* argument belonging to the predicted extension. Now, if we deviate from VAF and assume value as the only relevant independent variable, then we get more effective predictions, hovering at about 77% in C1 and 64% in C2 (average between both frames). This implies that, taking into account only the promoted values, the predictions improve by about 35 percentage points with respect to VAF. When both values are perceived with the same importance, the VAF-based prediction is just the same as that of Dung's model, which show success in predicting the selection of some argument in the extension is around 90% in C1 and 79% in C2 (both frames considered in each condition).

With respect to biases, some of the participants changed the degree of importance given to at least one value from one frame to the other, which hovers at about 14% in C1 ($n = 5$) and 20% in C2 ($n = 8$). In addition, part of the participants varied the order or the relative importance between the values from one frame to the other, though the changes mainly consisted in going from different degrees to equivalent or vice versa. In this respect, we observed very similar proportions of change in both conditions, around 20%. In C1, 71% (5 out of 7) of the variations were from equivalence in F1 to difference in F2, of which 60% ($n = 3$) changed in favor of evidence and 40% ($n = 2$) in favor of legality. In C2, 88% (7 out of 8) varied from difference in F2 to equivalence in F1, of which 29% ($n = 2$) changed in favor of evidence and 71% ($n = 5$) in favor of legality. This shows that individuals have a slight tendency to move towards legality in the context of F1 and towards evidence in the context of F2. In both cases, variations in the perception of the relative importance of the values tend to concede more strength to lenient arguments (X cannot be convicted, X was not responsible of the crime).

Regarding the question about any other value involved that influenced the choice of arguments, and which arguments supported that value, B and B′ were the most mentioned arguments (both conditions considered, 5 mentions in F1 and 7 in F2, respectively) and were associated with values such as responsibility, truth, justice and ethics. Next we have A and A′ (3 mentions in F1 and 4 in F2, resp.), identified with justice, and C and C′ (2 mentions in each frame), linked to law. In the vast majority of cases, influences were recognized as favoring the choice of those arguments, but not the rejection of others.

The results on intersubject argument acceptance are summarized in Table 9. The first and second columns show the percentages of acceptance in each frame and condition. Although the percentage variations could suggest some order effect that causes the acceptance of arguments in the second frame with different strength than in the first one, they are not statistically significant. This could be due to a small size of the sample. In any case, combining the results obtained in both conditions (third column)

Table 9

Variations of argument acceptance among conditions and frames in Experiment #4. Arrows represent the order in which frames were presented in each condition

C1 ($n = 35$)			C2 ($n = 39$)			Combined ($n = 74$)			
F1	→	F2	F1	←	F2	F1	F2	First frame	Second frame
A: 17%		A′: 29%	A: 31%		A′: 21%	A: 28%	A′: 24%	19%	30%
B: 43%		B′: 41%	B: 41%		B′: 41%	B: 42%	B′: 41%	42%	41%
C: 40%		C′: 32%	C: 38%		C′: 44%	C: 39%	C′: 38%	42%	36%

Table 10

Experiment #4. Accessions to give more importance to this value than to the other one

N = 74		Evidence	Legality	Equal
C1	F1	49%	11%	40%
	F2	51%	11%	37%
C2	F1	41%	13%	44%
	F2	56%	15%	28%
Average		49%	13%	37%

we obtained similar values between F1 and F2. Moreover, we still observed non-significant differences when combining the results in the first frame of each condition, that is, when participants answered without any previous exercise (last subcolumn), non-significant differences were still observed.[12]

Discussion. One important aspect of VAF is to model *objective* argument acceptance (Definition 4), as a way of capturing persuasion for any audience. In the frames involved in this experiment, according to the model, C and C′ are objectively acceptable in their respective frames: no argument defeats them, and if evidence is preferred to legality, then they are acceptable since they do not pose any threat to B and B′. On the contrary, results showed that individuals that prefer a value preferred to that perceived to be promoted by a given argument tend to make that argument not eligible. This result coincides with that of Experiment #3.

Now, a certain interpretation of Table 9 gives us a different picture regarding VAF-based predictions. In the last column, we can see that the arguments B (B′) and C (C′) have similar percentages of acceptance (around 40% on average), above those of A (A´) (around 26%). Taking into account that, on average, evidence is preferred over legality in 49% cases, against a 13% preference for legality over evidence (Table 10), we have a coincidence between the *collective acceptance* of those arguments and the outcome predicted by VAF. That is, if we ask which arguments are more likely to be collectively accepted in F1 and F2, since that evidence is at least as preferred as legality, then the preferred extensions {B, C} and {B′, C′}, respectively, give us the most probable acceptable arguments. Nevertheless, we do not have a good explanation, in terms of the model, about why this happens, because the data do not show the expected correlations according to VAF's semantical considerations. In fact, 66% of the participants that preferred evidence to legality did not accept C or C′ (on average).

Another possible explanation[13] is that participants do not actually see a conflict between arguments B and C (resp., B′ and C′). An explicit conflict could be created by adding "and therefore should be convicted" ("and therefore should be acquitted") at the end of B (B′). But then the conflict between B

[12]The comparisons between the frames on the acceptance of arguments by Fischer exact tests gave us the following values ($p < 0.05$): A-A′: 0.1794; B-B′: 1; C-C′: 0.6136.

[13]This was suggested by a reviewer.

and C is a bidirectional attack. It only becomes unidirectional in favor of C (C′) when legality is preferred over evidence.

In terms of framing, on the one hand, we observed in both conditions that, when moving from one frame to another with an isomorphic structure but contrary conclusions, individuals tend to modulate the strength of argument acceptance and rejection towards a more balanced assessment. However, the statistical differences are not significant, maybe due to the small sample size; hence, we plan to conduct more experiments in the future to address these issues. On the other hand, a tendency to change perceptions about the relative importance of the values involved in the argumentation frames were correlated to some extent with the perception of greater strength in lenient arguments. The data indicate that the value promoted by lenient arguments could be perceived as more important than the value promoted by harsh arguments. This is in line with findings about the influence of leniency bias on mock jury deliberations [39] and the outcome favorability as a strong determinant of individuals' willingness to accept authoritative decisions [32].

5. General discussion and concluding remarks

Since [45], the use of empirical methods borrowed from experimental psychology to test the formal semantics of argumentation frameworks has gained some popularity [23,25,30,36,41,46,47,49,50]. In this work we have taken a step further, and applied that methodology to test value-based argumentation frameworks.

The approach had several motivations. One of them arises from the very methodological limitations (basically, inherited from research in non-monotonic reasoning) of building semantics of argumentation frameworks on common intuitions about the solution of a handful of benchmark problems. Extension semantics are not formal semantics in a strict logical sense. In the latter, a semantics defines truth conditions of sentences and a notion of entailment, while extension semantics just "provides a way to select "reasonable" sets of arguments among all the possible ones, according to some criterion embedded in its definition" [5]. In consequence, researchers have adopted ideas of "soundness" just on the basis of the mentioned intuitions. Baroni and Giacomin [6] have also proposed a series of principles or properties with which to evaluate extension semantics, but these are only formalized expressions of the same intuitions. Hence, it could seem legitimate to "advocate the use of psychological experiments as a methodological tool for informing and validating intuitions about argumentation-based reasoning" [45]. In this regard, we agree more with "informing" than with "validating". For many years now, since the times of [51], cognitive psychology has been accumulating evidence that people deviate from formal, normative models of reasoning. However, although the experiments we reported here are along that line with respect to VAF, we cannot conclude that they undermine or invalidate its value as a normative model. VAFs can still be seen as idealizations of rational audiences that are able to ignore any unspoken, additional elements (values, biases, desires, etc.) in order to decide which arguments are better justified. Moreover, evidence can be used to generate new insights or modify old ones so that the model improves. Another motivation is to adopt a descriptive view and test VAF as a scientific theory to explain and predict human value-based argumentation. In this sense, contrasting with empirical evidence seems more legitimate as a means of validation. Our motivation, in sum, relies on using empirical data to, on the one hand, test VAF as a descriptive theory and, on the other, gain insights to improve it as a normative model. A clear limitation of our study is that we relied on very few scenarios, so we plan to explore more varied argumentative situations in the future to get greater conclusive force.

Experiments #1 and #2 allowed us to determine that people's argument acceptance deviates from the predictions based on VAF's semantics and is rather correlated with the importance given to the promoted values, regardless of the perceptions of argument interactions through attacks and defeats. On the other hand, most participants identified incompatibility between a defeated argument and its attacker, which seems to confirm the intuition that defeat presupposes conflict as a necessary condition. This has an interesting consequence regarding VAF. According to the model, if an attack is unsuccessful because the attacked argument promotes some preferred value, then the conflict between those arguments disappears, leading to accept possibly both arguments (see the notion of *conflict-freeness* in Definition 4, third item). However, the experiments revealed that, in such cases, people tend to reject the attacker, which seems to confirm that the incompatibility persists. Maybe this persisting incompatibility could be formally modeled as a reversed attack. This raises the question of whether the attacks should indeed be reduced to a symmetrical relation of opposition, resolved in one direction or another depending on how the preference of values determines the defeats. The problem of the directionality of the attacks was recurring throughout the experiments. Cramer and Guillaume [29] have conducted some cognitive experiments showing that people (both naive and experts on argumentation) judge attacks in correspondence with ASPIC+-based predictions. However, judgments were evaluated according to how individuals chose between pairs of arguments (accept, reject, or undecided status) and, in the case of naive participants, they were instructed on how to do that (e.g., they should not base their judgment on their knowledge; by default, an argument should be accepted unless the other argument provides reasons to reject it; etc.). This contrasts with our experiments, where participants did not receive any instruction. In any case, it is clear that the problem deserves further and deeper studies.

Our results also showed that each argument can be perceived as promoting more than one value with different degrees of relative importance. In [27], the authors investigated Bench-Capon and Sartor's case-based reasoning system [10] in that line, but the negative results on predicting good explanations for legal reasoning suggest the need for more research to fit the model. In the context of VAF, we thought of extending the model to allow representing arguments that promote more than one value by introducing a function $vals : AR \rightarrow 2^V$. Then, two alternative notions of *defeat* could be explored. One was to consider a comparison between sets of promoted values (e.g., [9,27,42]). Another one was to examine the degrees of promotion of values (e.g., [48]). We offered general expressions of those notions in Definition 6 and Definition 8, respectively. There is a clear link here with the argumentation schemes and critical questions approach to argument generation, as discussed in [1] and [18]. Although the representational facilities of the approach seem clear, its adequacy, both from a normative and descriptive point of view, should be studied in depth.

Experiments #3 and #4 evidenced that *objectivity*, understood as the acceptance of an argument for any audience (i.e., any order of the values), had no empirical correlation with the acceptance attitudes of participants. Experiment #3 is especially relevant because participants largely identified attacks and promoted values as expected but, again, preference over values tended to influence acceptance of arguments that promoted more preferred values and rejection of arguments that promoted less preferred values. This seems to confirm Dov Gabbay's opinion expressed in the following example: "If [politicians] are criticized by the Church or by some Nobel Prize winner economists, or by some experts, it is best to get another expert or another Nobel Prize winner or another bishop to support them!". In more general terms, to persuade individuals, it is better to get an argument promoting the value their prefer. In any case, our results are not conclusive and further investigation should include other scenarios, frameworks, and perhaps different types of questions.

Experiment #4 also presented some framing effects. Under similar structural factors, a percentage of the participants tended to vary the degrees of importance given to the values from one frame to the other (F1 and F2), while the arguments with the same position in the respective graphs were accepted with different strength. This may be due to the occurrence of a leniency bias towards the accused, according to the information from the framework. Some psychological findings could explain our results. For instance, McCoun and Kerr [39] showed that in mock juries, given two different decision procedures with various outcomes, such as convict or acquit, people tend to choose the procedure that leads to the benevolent outcome. Then, there could be a similar effect in the face of two structurally identical frameworks but with distinct framings and outcomes, such that people tend to choose different decision procedures (say, extension semantics) according to the bias. In the same way, Esaiasson et al. [32] argued that the tendency to prefer decisions leading to favorable outcomes is usually stronger than the preference for fair procedures. Mercier and Sperber [40], moreover, claimed that skilled arguers are not after the truth but after arguments supporting their views. The authors also argued that participants tend to show biased evaluations, analyzing the arguments contrary to their opinions, in which they look for flaws such as fallacies, and end up finding some. In addition to biases, people hardly evaluate the arguments only with the information offered, but take into account their own information and arguments.

In sum, persuasion depends to a large extent on psychological and informational factors, so the design of a normative model entails the arduous task of discerning which of these factors are in fact necessary for persuasion. In this vein, for example, Bench-Capon, Atkinson and McBurney [18] combined an action-based alternating transition system [1] with VAF to model some game-theory problems (particularly, the dictator and the ultimatum games), and their approach can account for framing effects described in the literature. Depending on the way a problem is described, different arguments are available, leading people to make distinct decisions even though the utility is the same in all frames. Hence, the behavior can be rationalized by analyzing the interaction of the arguments in the model. The work is a good example of how experimental research can provide information and insights to develop practical argumentation models.

Acknowledgements

We thank three anonymous reviewers for detailed comments and criticisms that notoriously improved the article. This research was financially supported by Agencia Nacional de Promociones Científicas y Tecnológicas (PICT 2017-1702) and Universidad Nacional del Sur (PGI 24/I265), Argentina.

References

[1] K. Atkinson and T. Bench-Capon, Practical reasoning as presumptive argumentation using action based alternating transition systems, *Artificial Intelligence* **171**(10) (2007), 855–874. doi:10.1016/j.artint.2007.04.009.

[2] K. Atkinson, T. Bench-Capon and P. McBurney, *Computational Representation of Practical Argument, Synthese* **152**(2) (2006), 157–206. doi:10.1007/s11229-005-3488-2.

[3] K. Atkinson and T.J. Bench-Capon, Value-based argumentation, *Journal of Applied Logics* **8**(6) (2021), 1543–1588.

[4] K. Atkinson and A. Wyner, The value of values: A survey of value-based computational argumentation, in: *From Knowledge Representation to Argumentation in AI, Law and Policy Making: A Festschrift in Honour of Trevor Bench-Capon on the Occasion of His 60th Birthday*, 2013.

[5] P. Baroni, M. Caminada and M. Giacomin, An introduction to argumentation semantics, *The Knowledge Engineering Review* **26**(4) (2011), 365–410. doi:10.1017/S0269888911000166.

[6] P. Baroni and M. Giacomin, On principle-based evaluation of extension-based argumentation semantics, *Artificial Intelligence* **171**(10) (2007), 675–700. doi:10.1016/j.artint.2007.04.004.

[7] P. Baroni, M. Giacomin and G. Guida, SCC-recursiveness: A general schema for argumentation semantics, *Artificial Intelligence* **168**(1–2) (2005), 162–210. doi:10.1016/j.artint.2005.05.006.
[8] T. Bench-Capon and H. Prakken, Using argument schemes for hypothetical reasoning in law, *Artificial Intelligence and Law* **18**(2) (2010), 153–174. doi:10.1007/s10506-010-9094-8.
[9] T. Bench-Capon, H. Prakken and W. Visser, Argument schemes for two-phase democratic deliberation, in: *Proceedings of the 13th International Conference on Artificial Intelligence and Law, ICAIL'11*, Association for Computing Machinery, New York, NY, USA, 2011, pp. 21–30. ISBN 9781450307550. doi:10.1145/2018358.2018361.
[10] T. Bench-Capon and G. Sartor, A model of legal reasoning with cases incorporating theories and values, *Artificial Intelligence* **150**(1) (2003), 97–143. doi:10.1016/S0004-3702(03)00108-5.
[11] T.J. Bench-Capon, Agreeing to differ: Modelling persuasive dialogue between parties with different values, *Informal Logic – Windsor Ontario* **22** (2002), 231–246.
[12] T.J. Bench-Capon, Representation of case law as an argumentation framework, in: *Proceedings of the Fifteenth Annual Conference on Legal Knowledge and Information Systems (Jurix 2002)*, IOS Press, 2002, pp. 102–112.
[13] T.J. Bench-Capon, Try to see it my way: Modelling persuasion in legal discourse, *Artificial Intelligence and Law* **11**(4) (2003), 271–287. doi:10.1023/B:ARTI.0000045997.45038.8f.
[14] T.J. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, *Journal of Logic and Computation* **13**(3) (2003), 429–448. doi:10.1093/logcom/13.3.429.
[15] T.J. Bench-Capon, Before and after dung: Argumentation in AI and law, *Argument & Computation* **11**(1–2) (2020), 221–238. doi:10.3233/AAC-190477.
[16] T.J. Bench-Capon, in: *Audiences and argument strength, The Third Workshop on Argument Strength*, Hagen, 11th to 13th October, 2021, Hagen, http://argstrength2021.argumentationcompetition.org/programme.html.
[17] T.J. Bench-Capon, K. Atkinson and A. Chorley, Persuasion and value in legal argument, *Journal of Logic and Computation* **15**(6) (2005), 1075–1097. doi:10.1093/logcom/exi058.
[18] T.J. Bench-Capon, K. Atkinson and P. McBurney, Using argumentation to model agent decision making in economic experiments, *Autonomous Agents and Multi – Agent Systems* **25**(1) (2012), 183–208. doi:10.1007/s10458-011-9173-6.
[19] T.J. Bench-Capon, S. Doutre and P.E. Dunne, Audiences in argumentation frameworks, *Artificial Intelligence* **171**(1) (2007), 42–71. doi:10.1016/j.artint.2006.10.013.
[20] T.J. Bench-Capon, H. Prakken and G. Sartor, Argumentation in legal reasoning, in: *Argumentation in Artificial Intelligence*, Springer, 2009, pp. 363–382. doi:10.1007/978-0-387-98197-0_18.
[21] F. Bex, K. Atkinson and T. Bench-Capon, Arguments as a new perspective on character motive in stories, *Literary and Linguistic Computing* **29**(4) (2014), 467–487. doi:10.1093/llc/fqu054.
[22] F. Bex and T. Bench-Capon, *Understanding Narratives with Argumentation, Computational Models of Argument*, IOS Press, 2014, pp. 11–18. doi:10.3233/978-1-61499-436-7-11.
[23] E. Bezou Vrakatseli, Empirical Evaluation of Simple Reinstatement in Formal Models of Argumentation, Master thesis, Utrecht University, 2021, Accepted: 2021-02-18T19:00:17Z.
[24] C. Cayrol and M.-C. Lagasquie-Schiex, Bipolar abstract argumentation systems, in: *Argumentation in Artificial Intelligence*, G. Simari and I. Rahwan, eds, Springer US, Boston, MA, 2009, pp. 65–84. doi:10.1007/978-0-387-98197-0_4.
[25] F. Cerutti, N. Tintarev and N. Oren, *Formal Arguments, Preferences, and Natural Language Interfaces to Humans: An Empirical Evaluation, ECAI 2014*, IOS Press, 2014, pp. 207–212. doi:10.3233/978-1-61499-419-0-207.
[26] F. Cerutti, N. Tintarev and N. Oren, Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation, in: *ECAI*, 2014, pp. 207–212.
[27] A. Chorley and T.J. Bench-Capon, An empirical investigation of reasoning with legal cases through theory construction and application, *Artificial Intelligence and Law* **13**(3) (2005), 323–371.
[28] J.L. Coleman, *Risks and Wrongs*, Oxford University Press, Oxford, 2002. ISBN 978-0-19-925361-6. doi:10.1093/acprof:oso/9780199253616.001.0001.
[29] M. Cramer and M. Guillaume, *Directionality of Attacks in Natural Language Argumentation, CEUR Workshop Proceedings*, RWTH Aachen University, 2018. https://orbilu.uni.lu/handle/10993/37027.
[30] M. Cramer and M. Guillaume, Empirical study on human evaluation of complex argumentation frameworks, in: *Logics in Artificial Intelligence*, F. Calimeri, N. Leone and M. Manna, eds, Springer International Publishing, Cham, 2019, pp. 102–115. doi:10.1007/978-3-030-19570-0_7.
[31] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* **77**(2) (1995), 321–357. doi:10.1016/0004-3702(94)00041-X.
[32] P. Esaiasson, M. Persson, M. Gilljam and T. Lindholm, Reconsidering the role of procedures for decision acceptance, *British Journal of Political Science* **49**(1) (2019), 291–314. doi:10.1017/S0007123416000508.
[33] D.M. Gabbay, Systems of interacting argumentation networks, *Journal of Logics and their Applications* **1**(1) (2014), 37–83.
[34] A.J. García and G.R. Simari, Defeasible logic programming: An argumentative approach, *Theory and Practice of Logic Programming* **4**(1–2) (2004), 95–138. doi:10.1017/S1471068403001674.

[35] Heuristics Biases, *The Psychology of Intuitive Judgment*, Cambridge University Press, 2002. doi:10.1017/CBO9780511808098.
[36] A. Hunter and M. Thimm, *Probabilistic Argument Graphs for Argumentation Lotteries, Computational Models of Argument*, IOS Press, 2014, pp. 313–324, https://ebooks.iospress.nl/doi/10.3233/978-1-61499-436-7-313.
[37] D. Kahneman and A. Tversky (eds), *Choices, Values, and Frames, Choices, Values, and Frames*, Vol. 840, Cambridge University Press, New York, NY, US, 2000, 840 pp. ISBN 978-0-521-62172-4, 978-0-521-62749-8.
[38] S. Lichtenstein and P. Slovic, Reversals of preference between bids and choices in gambling decisions, *Journal of Experimental Psychology* **89** (1971), 46–55. doi:10.1037/h0031207.
[39] R.J. MacCoun and N.L. Kerr, Asymmetric influence in mock jury deliberation: Jurors' bias for leniency, *Journal of Personality and Social Psychology* **54**(1) (1988), 21–33. doi:10.1037/0022-3514.54.1.21.
[40] H. Mercier and D. Sperber, Why do humans reason? Arguments for an argumentative theory, *The Behavioral and Brain Sciences* **34**(2) (2011), 57–74, ISSN 1469-1825. doi:10.1017/S0140525X10000968.
[41] S. Polberg and A. Hunter, Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches, *International Journal of Approximate Reasoning* **93** (2018), 487–543, https://www.sciencedirect.com/science/article/pii/S0888613X17303808. doi:10.1016/j.ijar.2017.11.009.
[42] H. Prakken, An exercise in formalising teleological case-based reasoning, *Artificial Intelligence and Law* **10**(1) (2002), 113–133. doi:10.1023/A:1019536206548.
[43] H. Prakken, An abstract framework for argumentation with structured arguments, in: *Argument & Computation*, Vol. 1, IOS Press, 2010, pp. 93–124, https://content.iospress.com/articles/argument-and-computation/456935. doi:10.1080/19462160903564592.
[44] H. Prakken and G. Sartor, Argument-based extended logic programming with defeasible priorities, *Journal of Applied Non-Classical Logics* **7**(1–2) (1997), 25–75. doi:10.1080/11663081.1997.10510900.
[45] I. Rahwan, M.I. Madakkatel, J.-F. Bonnefon, R.N. Awan and S. Abdallah, Behavioral experiments for assessing the abstract argumentation semantics of reinstatement, *Cognitive Science* **34**(8) (2010), 1483–1502. doi:10.1111/j.1551-6709.2010.01123.x.
[46] A. Rosenfeld and S. Kraus, Argumentation theory in the field: An empirical study of fundamental notions, in: *ArgNLP 2014, Frontiers and Connections Between Argumentation Theory and Natural Language Processing, Vol. 1341, CEUR Workshop Proceedings*, 2014, http://ceur-ws.org/Vol-1341.
[47] A. Rosenfeld and S. Kraus, Providing arguments in discussions on the basis of the prediction of human argumentative behavior, *ACM Trans. Interact. Intell. Syst.* **6**(4) (2016). doi:10.1145/2983925.
[48] G. Sartor, Doing justice to rights and values: Teleological reasoning and proportionality, *Artificial Intelligence and Law* **18**(2) (2010), 175–215. doi:10.1007/s10506-010-9095-7.
[49] M. Thimm, A probabilistic semantics for abstract argumentation, in: *Proceedings of the 20th European Conference on Artificial Intelligence, ECAI'12*, IOS Press, NLD, 2012, pp. 750–755, https://dl.acm.org/doi/abs/10.5555/3007337.3007468. ISBN 9781614990970.
[50] A. Toniolo, T. Norman and N. Oren, Enumerating preferred extensions: A case study of human reasoning, in: *Theory and Applications of Formal Argumentation*, E. Black, S. Modgil and N. Oren, eds, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, Germany, 2018, pp. 192–210. doi:10.1007/978-3-319-75553-3_14.
[51] A. Tversky and D. Kahneman, The framing of decisions and the psychology of choice, *Science* **211**(4481) (1981), 453–458. doi:10.1126/science.7455683.