Argument & Computation -1 (2024) 1–25
DOI 10.3233/AAC-230016
IOS Press
CORRECTED PROOF

Experimental evaluation of gradual argument acceptability semantics: The case of reinstatement

Elfia Bezou Vrakatseli [a], Henry Prakken [b,*] and Christian P. Janssen [c]
[a] *Department of Informatics, King's College London, United Kingdom*
E-mail: elfia.bezou_vrakatseli@kcl.ac.uk
[b] *Department of Information and Computing Sciences, Utrecht University, The Netherlands*
E-mail: h.prakken@uu.nl
[c] *Experimental Psychology and Helmholtz Institute, Utrecht University, The Netherlands*
E-mail: c.p.janssen@uu.nl

Abstract. This paper investigates whether empirical findings on how humans evaluate arguments in reinstatement cases support the 'fewer attackers is better' principle, incorporated in many current gradual notions of argument acceptability. Through three variations of an experiment, we find that (1) earlier findings that reinstated arguments are rated lower than when presented alone are replicated, (2) ratings at the reinstated stage are similar if all arguments are presented at once, compared to sequentially, and (3) ratings are overall higher if participants are provided with the relevant theory, while still instantiating imperfect reinstatement. We conclude that these findings could at best support a more specific principle 'being unattacked is better than attacked', but alternative explanations cannot yet be ruled out. More generally, we highlight the danger that experimenters in reasoning experiments interpret examples differently from humans. Finally, we argue that more justification is needed on why, and how, empirical findings on how humans argue can be relevant for normative models of argumentation.

Keywords: Empirical evaluation, simple reinstatement, imperfect reinstatement, suspension of disbelief, graded acceptability

1. Introduction

Rahwan et al. [35] presented an empirical study of how people evaluate arguments in the context of counterarguments. Their aim was to assess how the abstract argumentation semantics of Dung [13] treat so-called reinstatement patterns, in which an argument that is attacked by another argument is defended, or 'reinstated', by an argument attacking the attacker, so that if there are no further arguments, the first and third argument are acceptable but the second argument must be rejected. They found that people by-and-large assess arguments according to Dung's semantics but not fully: on a 7-point scale, the first argument was rated significantly more acceptable when presented on its own than when presented together with its attacker and defender.

There are several reasons to reconsider these experiments. A general reason is that it has been claimed that the psychological sciences face a 'replicability crisis' since the results of many well-known experiments appear not to be replicable [28]. In light of this, one aim of this paper is to test whether the results

*Corresponding author. E-mail: h.prakken@uu.nl.

of Rahwan et al. [35] can be replicated. A more specific reason is that since the study of Rahwan et al. appeared, the study of gradual notions of argument acceptability has become popular. These studies include probabilistic [23], ranking-based [2], and graded [20] approaches. Probabilistic approaches assign probabilities to arguments, which either express the probability that an argument is part of an argumentation framework (Hunter & Thimm's [23] 'constellation approach') or the probability that an argument is acceptable (Hunter & Thimm's [23] 'epistemic approach'). Ranking-based approaches define a preference relation (which can be partial) on a set of arguments, to express that one argument is at least as acceptable as another one. Finally, graded approaches assign a number to arguments expressing their numerical strength. Such graded approaches propose semantics for computing the numbers or rankings and it proposes principles, or postulates, for these semantics. Some of the research on ranking-based and graded semantics refers to Rahwan et al.'s study for support of their approaches, either for gradual notions of acceptability in general [22,31] or for specific features of the new semantics [1,19,20]. In particular, Grossi and Modgil [19] cite Rahwan et al. in support for a principle that everything else being equal, having fewer attackers is better. This principle is also a key element in several of the new semantics. For instance, all six ranking-based semantics studied by Bonzon et al. [4] satisfy the principle of 'void precedence' [2], according to which an argument that has no attackers is more acceptable than an argument that has attackers, even if these attackers are counterattacked.

Accordingly, another aim of this paper is to investigate whether Rahwan et al.'s study indeed provides support for these recent developments, in particular for the void precedence principle that was the focus of their experiments. In doing so, we will regard these formalisms not as descriptive but as prescriptive, or normative models of argumentation, that is, as modelling how people *should argue*. Our investigations are in part motivated by discussions of Cramer and Guillaume [10,11,21] and Prakken and de Winter [34] of Rahwan et al.'s study, which give reasons to be cautious when referring to Rahwan et al. in support of the new semantics, suggesting alternative explanations for Rahwan et al.'s findings. In doing so, we do not aim to question the importance of graduality in argumentation as such. We take it for granted that graduality plays important roles in argument evaluation; the question that concerns us is how these roles can best be modelled. Moreover, we would also like to note that not all graduality semantics regard the void precedence principle as generally acceptable; for example, Bonzon et al. [5] and Thimm and Kern-Isberner [38] independently challenge the principle for separate reasons.

In this paper, we report on three experiments in which humans evaluate arguments. The first experiment succeeded in replicating Rahwan et al.'s results on imperfect recovery from attack. The other two were aimed to test two versions of an alternative explanation for Rahwan et al.'s results suggested by Rahwan et al. and Prakken and de Winter [34], namely, that the imperfect recovery of arguments from attack is not because the participants in the experiments applied the void precedence principle when rating the arguments, but it is due to the specific way in which the arguments were presented to them. These experiments yielded mixed results. We evaluate the results of our experiments in light of the above-mentioned literature but also in light of the question whether empirical studies have anything to say at all about the assessment of normative theories of argumentation. Our main conclusion will be that the results of Rahwan et al. [35] cannot (yet) be considered supporting evidence for the idea that – all other things being equal – having fewer attackers is better, not even for its special case, the 'void precedence' principle, since alternative explanations for the effect they found cannot be ruled out and since a more convincing explanation is needed for why empirical findings are relevant for normative theories of argumentation.

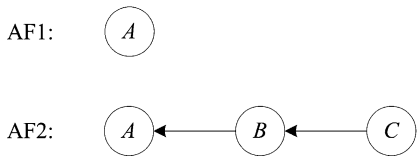

Fig. 1. The reinstatement pattern.

2. Preliminaries

In this section, the basics of Dung's theory of abstract argumentation frameworks are summarised and applied to the reinstatement pattern that was the subject of the studies of Rahwan et al. [35]. We present Dung's semantics in a labelling version, which is equivalent to Dung's original semantics [7,25].

An *abstract argumentation framework* (*AF*) is a pair $\langle \mathcal{A}, \mathcal{C} \rangle$, where $\mathcal{A}$ is a set of *arguments* and $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation of *attack*. The labelling approach characterises the various semantics in terms of labellings of $\mathcal{A}$.

A *labelling* of an abstract argumentation framework $\langle \mathcal{A}, \mathcal{C} \rangle$ is any assignment of exactly one of the labels *in*, *out* or *undecided* to all arguments from $\mathcal{A}$ such that:

(1) an argument is *in* iff all arguments attacking it are *out*.
(2) an argument is *out* iff it is attacked by an argument that is *in*.
(3) an argument is *undecided* iff it is neither *in* nor *out*.

Then *stable semantics* labels all arguments as either *in* or *out*, while *grounded semantics* minimises and *preferred semantics* maximises the set of arguments that are labelled *in*. Relative to a semantics, an argument is *sceptically acceptable* if it is labelled *in* in all labellings, it is *rejected* if it is labelled *out* in all labellings, and it is *credulously acceptable* if it is labelled *in* in at least one labelling.

The reinstatement pattern studied by Rahwan et al. is displayed in Fig. 1. In both *AF*s argument *A* is sceptically acceptable in all three semantics. With only *A* this is trivial since *A* has no attackers. When also *B* and *C* are present, *C* has to be made *in* by constraint (1), since it has no attackers, and *B* has to be made *out* by constraint (2), thus *A* has to be made *in* by constraint (1). Thus *C* reinstates *A* by defending *A* against *B*.

This outcome for *AF*2 is the same if the attack from *B* on *A* is made symmetric but it changes if the attack from *C* on *B* is made symmetric (regardless whether the same is done for *B*'s attack on *A*). If *C* and *B* attack each other then the just-given labelling is still possible but it is not the only one: a labelling in which *B* is *in* and both *A* and *C* are *out* also satisfies the constraints. Both of these labellings are preferred and stable but not grounded, since the empty labelling also satisfies the constraints. Thus all three arguments are credulously acceptable in preferred and stable semantics while they are not acceptable in grounded semantics.

Rahwan et al. presented six examples to the participants in their experiments, all having the same pattern and all assumed to instantiate *AF*2 from Fig. 1 (all the examples can be found in Appendix C). The participants were first confronted with a single argument, for instance:

A: *The battery of Alex's car is not working. Therefore, Alex's car will halt.*

They were then asked to rate their confidence in its conclusion. Only then were they subsequently confronted with an attacker and defender, for instance:

B: *The battery of Alex's car has just been changed today. Therefore, the battery of Alex's car is working.*
C: *The garage was closed today. Therefore, the battery of Alex's car has not been changed today.*

After both arguments, the participants were again asked to rate their confidence in the conclusion of the initial argument. After argument B, their average rating of A's conclusion went down while after argument C was presented to them, their average rating went up again, but to a significantly lower level than after being presented with A only. Rahwan et al. concluded that their results support the notion of reinstatement but not fully, since a reinstatement argument does not fully recover from an attack.

One explanation Rahwan et al. consider for their result is in terms of an effect of 'suspension of disbelief', according to which participants are capable of thinking of different kinds of objections to the presented arguments but they suspend these objections for the sake of the experiment. However, when one objection is presented by the experimenter, this suspension is disrupted and some participants start to let their private beliefs 'leak' into their assessments of the arguments. Prakken and de Winter [34] suggest a variation of this explanation, advocating that after being introduced to an attacker, a participant's degree of belief in other possible attackers increases as well since the very introduction of an attacker leads them to consider other possible objections.

3. The experiments

We conducted three experiments to investigate whether empirical findings on how humans evaluate arguments in reinstatement cases support the 'void precedence' principle. The methods of the experiments overlap and are presented together for brevity. Experiment 1 is an online replication of the study by Rahwan et al. [35]. Specifically, we test whether rating is lower at the reinstated stage compared to the base case when arguments are presented one-by-one (cf. Rahwan et al.). Based on this replication, we then test ideas proposed by Rahwan et al. [35] and Prakken and de Winter [34]. Specifically, experiment 2 tests whether the rating is different if all arguments (including the attack and the defense) are presented at once. Finally, experiment 3 tests what happens if first all possible scenarios are presented – i.e., generalised forms of the arguments the participants encounter during evaluation[1] – and then the arguments are presented one-by-one. As an example of (3), the generalised form of the car battery example was:

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

3.1. Hypotheses

We tested the following four hypotheses.

Hypothesis 1: When arguments are presented sequentially (experiment 1), participants' ratings for the conclusion of argument A in the reinstated stage are lower than in the base stage but higher than after argument B is presented.

The first hypothesis merely predicts a successful replication of Rahwan et al.'s results. Note that our participant number (130 aimed) is significantly higher than that used by Rahwan et al. (20), to gain further confidence in the result.

[1] All generalized forms can be found in Appendix C.

Hypothesis 2: When all arguments are presented at once (experiment 2), participants' ratings for the conclusion of argument *A* are higher than the (corresponding) ratings in the reinstated stage of the first case/manner-of-presentation (where all arguments are also available but have been introduced sequentially).

The second hypothesis suggests that when all the information is presented at the same time to the participants, the confidence in the conclusion of argument *A* is higher than the corresponding confidence in the reinstated stage when arguments have been presented one-by-one. Since the introduction of an attacker may change the participant's belief in the initially presented argument even after it has been reinstated, it is possible that it is the very gradual process of presentation that influences the participant's degree of belief. To quote Rahwan et al., "[p]articipants can easily generate all sorts of objections to the arguments presented to them by the experimenter, but they suspend their disbelief in these arguments for the sake of the experiment. When one objection is presented by the experimenter herself, though, suspension of disbelief is disrupted". Thus, if we eliminate the gradual factor of presentation, the initial suspension of disbelief may remain, since there is no stage where a *new* objection is presented that can disrupt it.

Possibly, when an attacker is introduced after one has placed their confidence in an argument, a kind of 'breach of confidence' is generated, one that cannot be later eradicated (by introducing another attacker) and that has caused the disruption of the initial experiment's 'convention/contract' (i.e., the suspension of disbelief). Hence, if all arguments were presented at once, they could all be considered as the aforementioned 'arguments presented by the experimenter' and participants would suspend their disbeliefs for all of them (as suggested). Provided with all the information (i.e., all the arguments in play) at the beginning, participants can make a deliberation without the element of surprise, resulting in giving the conclusion of argument *A* a higher confidence rating than in the reinstated stage of a gradual presentation.

Hypotheses 3a+3b: When participants are first presented with all possible scenarios (experiment 3) – i.e., when they are presented with generalised forms of the arguments they will encounter during evaluation, before evaluating them – and are then asked to evaluate the arguments one-by-one (the same way as in experiment 1):

a their ratings for the conclusion of argument *A* in the reinstated stage are higher than the corresponding ratings in the reinstated stage of the first experiment (where participants have not seen all the possible scenarios beforehand);
b their ratings for the conclusion of argument *A* in the base stage are lower than the corresponding ratings in the base stage of the first experiment.

In our statistical test, we ran an Analysis Of Variance (ANOVA) with experiment (experiment 1 or 3) as between-subjects factor, and moment (base stage versus reinstated stage) as within-subjects factor. Based on the hypotheses above, we would expect a significant interaction effect: rating is lower in the reinstated stage for participants in experiment 1 (compared to its base stage), whereas this is not the case for experiment 3 (i.e., no imperfect reinstatement is expected in experiment 3).

Extending our thinking concerning the second hypothesis, hypotheses 3a and 3b examine another possible explanation via a different manner of argument presentation. When a participant initially evaluates an argument, no evidence for or against its premises, inference, or conclusion has been offered, whereas after being presented with the attacker and defender, further evidence is overall provided, allowing the subject to form a more complete image of a precise situation.

Hypotheses 3a and 3b are based on [34], who argue that the introduction of an attacker increases the participants' degree of belief in other possible attackers, which are not explicitly ruled out in the presented arguments. They suggest that the introduction of a relevant theory prior to participants' evaluations will cause the confidence degree in the conclusion of argument A in the base stage to decrease (compared to ratings from the first manner of presentation) and to increase in the reinstated stage. Their suggestion is based on the assumption that if a participant was aware from the beginning of (all) the reasons argument A can be vulnerable, their belief in the possibility of the attacker that is presented (here, argument B) would increase from the base stage, resulting in a lower rating for the conclusion A at that stage. By the same logic, their degree of belief in all other attackers, which are not ruled out (but neither presented) in the experiment, would have no reason to increase after the actual introduction of the attacker in the defeated stage (contrarily to when one is not initially introduced to the whole theory) and, thus, confidence in argument A's conclusion would increase in the reinstated stage.

A confirmation of hypotheses 2, 3a, and 3b would underline the importance of the way in which subjects are presented with arguments, proving it affects participants' confidence. Such confirmations would support the observations of Rahwan et al. [35] and Prakken and de Winter [34] on the possible effects of suspension of disbelief, as, then, said findings could be interpreted as a result of the two aforementioned suggested explanations and not as support for graded notions of argument acceptability.

3.2. Method

We conducted three experiments. In all three experiments, participants had to evaluate the acceptability status of natural language arguments, in which we followed the method of Rahwan et al. [35] as closely as possible in terms of materials, procedure, and measurement, discussed in more detail below.

Participants. In each experiment, 130 participants took part (390 total). All were 18–65 years old. The average age was comparable between experiments (mean age for experiment 1, 2, and 3 respectively: 30, 33, and 28 years of age). All participants were volunteers, recruited through personal contact, and had no pre-knowledge of the topic of study. The participants were recruited from the general public, but the majority consisted of university students (for the remainder, no clear groups were identifiable). Participants were required to be over 18 years of age, and able to read and speak English, for which we probed them at the start of the survey.[2] All participants met the age and language requirements. The three samples were independent, meaning that each participant participated in only one experiment.

Materials. The materials followed original stimuli of Rahwan et al. as close as possible. In each experiment, participants had to rate eight sets of arguments, consisting of three arguments each, where the conclusion of each next argument contradicts a premise of the preceding argument. The first six sets were taken from Rahwan et al. while the two remaining sets were added by us in a similar style. [3] Specifically, these were:

A: *The power is out, so Claire cannot charge her phone.*
B: *The TV is playing, so the power is not out.*
C: *The TV is broken, so the TV is not playing.*

and

[2] The questionnaire that was used can be found in Appendix B.

[3] A list of all argument sets can be found in Appendix C, while an example of the way they were presented can be found in Appendix D.

A: *Animals have the right to be left unharmed, so we should ban animal testing.*
B: *Animals are very dissimilar to humans, so animals do not have such a right.*
C: *Animals resemble us anatomically, physiologically, and behaviourally (e.g., recoiling from pain, fearing tormentors), therefore they are not very dissimilar to humans.*

At various points (see Design), participants had to rate the acceptability of the conclusion of argument *A*. The ratings were given on a 7-point scale incrementally numbered from *Certainly false* (1) to *Certainly true* (7) as in Rahwan et al. [35].

Design. In experiment 1, we replicate Rahwan et al. [35]. Arguments *A*, *B*, and *C* were added in sequence. After each added statement, participants had to rate the acceptability of the conclusion of argument *A*. Consistent with hypothesis 1, and Rahwan et al. [35], we expect ratings to be higher after the presentation of argument *A* (base stage) compared to after the presentation of argument *C* (reinstated stage). This is tested with a paired t-test.

In experiment 2, all arguments are presented at once, and participants only provide one rating. We test whether this rating is different from the ratings at reinstated stage of experiment 1. Conform hypothesis 2, we expect ratings to be higher for participants from experiment 2.

In experiment 3, for each set of arguments, participants first received a text that included generalisations of all three arguments (an example of which can be found at the beginning of Section 3). They then had to rate the conclusion of argument *A* in a similar fashion as in experiment 1. As we now have a measurement at base and at reinstated stage for experiments 1 and 3, we analyse the results using an analysis of variance with experiment as between-subjects factor, and moment (base versus reinstated stage) as within-subjects factor. Conform hypothesis 3, we expect a significant interaction effect: in experiment 1 rating is expected to be lower in the reinstated stage; in experiment 3 we expected there to be no or little difference between the reinstated and the base stage.

Procedure. Participants did the experiment online using a Qualtrics (https://www.qualtrics.com/) survey. The participants were asked to complete the survey whenever they liked but were instructed to do so in a quiet environment without interruptions. In the survey, they were first asked a brief set of questions about their age and language capability. They then received a brief explanation of the study. Participants were then asked to rate four sets of arguments. The nature of questioning depended on which experiment they took part in (1, 2, or 3, see Design). Although we had 8 sets of arguments, each participant only rated 4 sets (randomised across participants).

Analysis. We removed data from participants whose response set was not complete (27, 34, and 20 participants in experiments 1, 2, and 3 respectively). We then calculated the average score for each rating type (reinstated stage, and base stage for experiments 1 and 3). In statistical analysis, we use alpha at .05 for significance.

3.3. Results

3.3.1. Experiment 1 and hypothesis 1

First, we test if our replication finds the same pattern of effect as Rahwan et al. [35]. Since one group of participants experienced both conditions (i.e., base and reinstated), we ran a paired t-test (i.e., per participant one can look at a pair of results).[4] The paired t-test on the data of experiment 1 found that

[4]Generally, a t-test compares whether two conditions (or groups) differ significantly from each other, and helps to make a yes/no decision about there being a significant difference between conditions. A paired t-test can be run in situations where one

Table 1
Experiment 1: ratings of group 1 per stage & set

Set	Mean per stage		
	Base	Defeated	Reinstated
1	5.66 ± 1.71	3.51 ± 2.14	5.42 ± 1.74
2	5.64 ± 1.56	3.76 ± 2.06	5.92 ± 1.18
3	4.71 ± 1.32	3.40 ± 1.49	4.25 ± 1.53
4	5.40 ± 1.38	3.72 ± 1.71	5.36 ± 1.56
5	6.12 ± 1.08	4.30 ± 1.71	4.98 ± 1.41
6	6.25 ± 1.09	3.85 ± 2.02	5.43 ± 1.67
7	5.69 ± 1.50	3.27 ± 1.79	4.86 ± 1.63
8	5.44 ± 1.55	4.18 ± 2.08	5.48 ± 1.58
Overall	5.61 ± 0.99	3.75 ± 1.08	5.21 ± 0.96

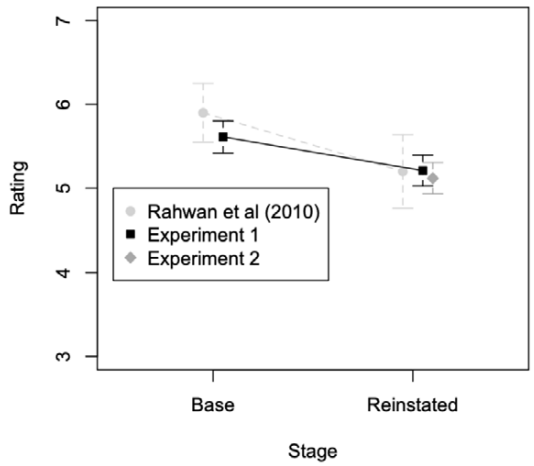

Fig. 2. Rating at base and reinstated for 2 experiments and Rahwan et al. (2010). Error bars show 95% confidence intervals; points are horizontally plotted slightly to the side of each other for better readability.

ratings at the base stage ($M = 5.61$, $SD = 0.99$, 95% $CI = [5.42, 5.81]$) were significantly higher compared to the reinstated stage ($M = 5.21$, $SD = 0.96$, 95% $CI = [5.02, 5.40]$), $t(102) = 4.636$, $p < .001$. Thus, ratings of argument A's conclusion are found to decrease after attacker B (see column 'Defeated' in Table 1) and increase again after counterattacker C, but not to the initial level (see column 'Reinstated' in the same table). This pattern is consistent with the original experiment [35]. Figure 2 shows this result and also presents the values observed in Rahwan et al. [35], while Table 1 summarises the ratings for each stage and set. The mean and standard deviation for each set are calculated from within-set ratings. The overall statistics are derived from the average ratings of each participant across all sets they evaluated. The number of the sets corresponds to the numbers of Appendix C, while the presentation of each stage can be found in Appendix D. It can be seen that, apart from the significant difference between conditions/stages, the observed values are also comparable between our study and

group of participants experienced both conditions; per participant one can look at the pair of results (i.e., whether ratings by the same participants are different after hearing argument A, versus after argument C). For further, technical, information on the t-test, see Field [16]. For a broader understanding of the rationale behind using statistical methods, see for example Cairns [6].

Table 2
Experiment 2: ratings of group 2 per set

Set	Mean
1	5.04 ± 1.67
2	5.70 ± 1.72
3	4.96 ± 1.66
4	5.10 ± 1.59
5	5.46 ± 1.27
6	5.04 ± 1.75
7	4.28 ± 1.60
8	5.31 ± 1.56
Overall	5.12 ± 0.93

Rahwan et al. [35] (specifically: in Fig. 2, there is a strong overlap between the error bars; the means of the two studies fall inside the region defined by the error bars). This confirms the first hypothesis and replicates the result of Rahwan et al. [35], this time with a considerably larger set of participants.

3.3.2. Experiment 2 and hypothesis 2

Next, we test if participants give higher ratings if information is presented all at once (experiment 2) compared to sequentially (experiment 1). As the groups had unequal numbers of participants, we ran an independent Welch t-test.[5] There was no significant effect of presentation manner on rating, $t(196.56) = -0.683$, $p = .496$. Thus, presenting all arguments at once before asking a rating of argument A's conclusion does not lead to higher ratings and, so, the second hypothesis cannot be confirmed. Indeed, Fig. 2 shows that the ratings in experiment 2 ($M = 5.12$, $SD = 0.93$) are similar (i.e., means are close, error bars overlap largely when the data of experiment 2 is compared to that of experiment 1 and of [35]). The ratings of each set can be found in Table 2, with the number of the sets corresponding to the numbers of Appendix C again. Following the same procedure as outlined for experiment 1, the mean and standard deviation for each set are calculated from within-set ratings, and the overall statistics are derived from the average ratings of each participant across all sets they evaluated.

3.3.3. Experiment 3 and hypothesis 3a and 3b

Next, we test if it makes a difference if participants are provided with generalisations of all three arguments first. To this end, after checking the equality of variances of each group/experiment with Levene's test, we ran a 2 (experiment: 1 or 3) x 2 (stage: base versus reinstated) mixed ANOVA.[6] We found a significant effect of experiment, $F(1, 211) = 12.906$, $p < .001$. There was also a significant effect of stage, $F(1, 211) = 53.66$, $p < .001$. There was no interaction between study and stage, $F(1, 211) = 1.227$,

[5] By contrast to our first experiment, when data is compared between two different groups of people, the data cannot be 'paired' and therefore an alternative test has to be used. This is the case for our comparison between experiment 1 and 2. As the experiments have different numbers of participants (i.e., unequal groups), we use a Welch t-test, which is statistically more appropriate for cases with unequal groups, to the end of comparing again whether there is a significant difference between conditions. For further, technical, information on the Welch t-test, see Field [17].

[6] ANOVA is used to compare the means across more than two groups or conditions to determine if there is a statistically significant difference among them. In our experiments, we used a mixed ANOVA to analyse the effects of different stages (Base vs. Reinstated) and experiments (Experiment 1 vs. Experiment 3) on participant ratings. The mixed ANOVA allows us to evaluate not only the main effects of each independent variable (such as 'experiment' or 'stage') but also their interaction effects, which can reveal whether the impact of one variable depends on the level of another variable. For instance, it helps us understand if changes in ratings between the base and reinstated stages differ significantly between different experiments. For further technical information on ANOVA, see Field [15].

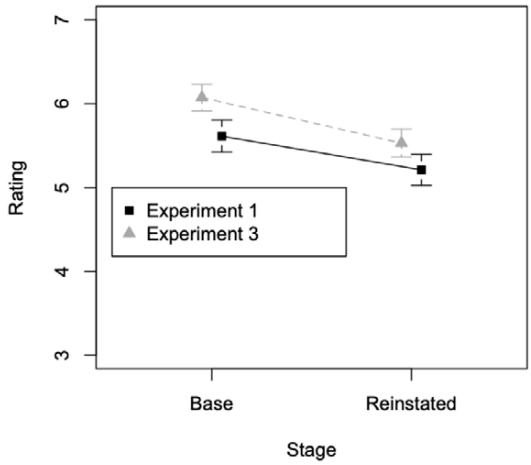

Fig. 3. Rating at base and reinstated for experiment 1 and 3. Error bars show 95% confidence intervals; points are horizontally plotted slightly to the side of each other for better readability.

$p = .269$. Figure 3 illustrates these effects. The parallel lines suggest that in both experiment 1 and experiment 3 ratings are higher in the base stage compared to the reinstated stage, and the reduction in rating between the two is comparable (i.e., main effect of stage). In addition, ratings in experiment 3 were higher in general (i.e., main effect of experiment). In other words, when participants first see the possible scenarios and then rate the arguments one-by-one, they rate A's conclusion higher in both the reinstated and base stage (compared to the corresponding stages of experiment 1). Thus, hypothesis 3a is confirmed but hypothesis 3b is rejected. This is contrary to our expectation of an interaction effect (i.e., we expected crossing lines in Fig. 3, with the line for experiment 3 being relatively flat). The expectation was that for experiment 3 the ratings in the reinstated stage were higher than those of experiment 1 (hypothesis 3a), but that in the base stage participants from experiment 3 provided a lower rating than those in experiment 1 (hypothesis 3b). We did not observe this interaction, as hypothesis 3a was confirmed but hypothesis 3b was rejected. The ratings for each set and stage are summarised in Table 3, with the mean and standard deviation for each set calculated from within-set ratings, and overall statistics derived from participants' average ratings across all sets (the numbers of the sets correspond to the numbers of Appendix C).[7]

4. Discussion

This study purported to (1) replicate the findings of Rahwan et al. [35] and (2) investigate whether these findings support the 'fewer attackers is better' principle, in particular its special case of the void precedence principle, incorporated in many current graded notions of argument acceptability, or whether alternative explanations suggested by Rahwan et al. [35] and Prakken and de Winter [34] undercut such support. To summarise our results, our experiment found that participants' ratings of argument A's conclusion decrease after seeing attacker B and increase again after seeing counterattacker C, but not to the initial level. This confirms our hypothesis 1 and replicates [35]'s findings. This is an important result since replicability is one of the cornerstones of the scientific method and since, as we noted in the

[7] More detailed graphs of the data for all three experiments can be found in Appendix A.

Table 3
Experiment 3: ratings of group 3 per stage & set

Set	Mean per stage		
	Base	Defeated	Reinstated
1	6.30 ± 1.25	4.73 ± 2.01	5.77 ± 1.66
2	6.09 ± 1.51	3.72 ± 2.01	6.08 ± 1.66
3	5.77 ± 1.53	4.13 ± 1.75	4.98 ± 1.70
4	5.35 ± 1.92	4.00 ± 1.97	5.52 ± 1.61
5	6.68 ± 0.51	4.07 ± 2.02	5.58 ± 1.31
6	6.29 ± 1.43	3.71 ± 2.33	5.65 ± 1.56
7	6.61 ± 1.17	3.63 ± 2.30	4.96 ± 1.88
8	5.57 ± 1.60	4.38 ± 1.95	5.66 ± 1.27
Overall	6.07 ± 0.85	4.06 ± 1.14	5.53 ± 0.89

introduction, social psychology is currently facing a replication crisis. In experiment 2, we found that presenting all arguments at once before asking a rating of argument A's conclusion did not lead to higher ratings compared to those observed in the sequential study of experiment 1 (rejecting hypothesis 2). In experiment 3, we found the opposite when the participants first see the possible scenarios and then rate the arguments after seeing the arguments one-by-one (confirming hypothesis 3a). Finally, in experiment 3 we found that participants rate A's conclusion higher in the base stage as well, compared to the base stage of experiment 1 (rejecting hypothesis 3b). Thus, we did not find the interaction effect that the confirmation of both hypotheses would entail. We now discuss various issues relevant to the question of whether our results strengthen the arguments for the 'fewer attackers is better' principle.

4.1. Generalisation to other patterns

We first recall an observation of Rahwan et al. [34] that, even if the results support a principle that 'an argument is better if it is unattacked than if it is attacked' in examples following the pattern of Fig. 1, the findings cannot be used as support for the more general intuition formalised in Grossi and Modgil's 'fewer attackers is better' principle [19,20] and Amgoud and Ben-Naim's void precedence principle [2], which intuition is at the heart of many current gradual and ranking-based approaches. The point is that the more general intuition also applies to structures where, unlike in Fig. 1, arguments A in $AF1$ and $AF2$ refer to different arguments. Neither in Rahwan et al.'s nor in our experiments examples of this kind were shown to the participants. So, at best Rahwan et al.'s and our experiments confirm the special case of the void precedence principle in which the arguments A in both AFs in Fig. 1 are the same argument. Note also that another kind of situation not studied by Rahwan et al. or us is when arguments have multiple and various numbers of attackers.

4.2. Suspension of disbelief

We next note that our results cast some doubts on Rahwan et al.'s suggested explanation in terms of suspension of disbelief and its variant suggested by Prakken and de Winter [34]. Rahwan et al. do not claim that the introduction of an attacker makes the subjects think/come up with objection, but rather that it causes them to disrupt their suppressing of their already existent objections. In this study, we hypothesised that if confronted with all three arguments at the same time, participants would apply their suspension of disbelief to all the (initially) presented arguments. As our hypothesis 2 is rejected –

i.e., introducing all three arguments at the same time does not have a significant effect on the subjects' confidence in A's conclusion – Rahwan et al.'s explanation regarding the disruption of suspension of disbelief cannot be validated.

The same holds for Prakken and de Winter's variant of the explanation in terms of suspension of disbelief [34], according to which the initial introduction of the relevant theory would have made the participants in group 3 aware of possible counterarguments from the start, unlike the participants in group 1. This should have led to the ratings for the conclusion of argument A in the base stage of group 3 being significantly lower than those of group 1, which was our hypothesis 3b. However, this hypothesis was rejected and, surprisingly, not only are the ratings of the third group not lower in the base stage, but they are actually significantly higher. Thus, this is a case where the possibility of an attacker was present from the beginning without it influencing negatively the ratings of the argument that could be attacked. The absence of the expected interaction effect suggests that – despite the introduction of the relevant theory beforehand – the recovery was not complete in the third group either and, thus, Prakken and de Winter's suggestion cannot explain imperfect reinstatement.

What is puzzling is the confirmation of hypothesis 3a in contrast to the rejection of hypothesis 3b, as what we expected was that the introduction of the theory would have opposite results on the base and reinstated stage. One reason why the introduction of the corresponding theory results in an increase of the ratings' level in both stages could be that when introduced with a theory beforehand, the participant gains reassurance. Even though aware of the possibility of an attacker, when an argument is unattacked, the participant has no reason/evidence not to believe it. Thus, the introduction of a *possible* attacker might in this case strengthen the attacker's *absence in the base stage*, thus increasing confidence in the conclusion of argument A. This could even be extended to the reinstated stage: participants might feel more reassured after being presented with the instantiation of the possibilities they were originally introduced with. This could also explain why a similar effect did not appear in the second group; in the third group, a participant is originally introduced to possibilities, which are later realised, whereas in the second group a participant misses this intermediate step of reassurance. However, the results of the second group could also be explained by the task of group 2, as we will further discuss in Section 4.4.

4.3. Natural language versus formalisation

At this point, it might be thought that our findings strengthen the support for the void precedence principle. The underlying idea here would be that the participants rated the arguments' conclusions with this principle in mind. We first discuss whether this explanation can be accepted on the basis of Rahwan et al. [35] and our experiments. Later, we will discuss to which extent such empirical claims and explanations are relevant for assessing normative models of argumentation.

There is yet another alternative explanation of the results, independently suggested by Prakken and de Winter [34] and Cramer and Guillaume [10,11,21], namely, that when rating the arguments, the participants may not have had the reinstatement pattern of Fig. 1 in mind but a different pattern. All argument sets in the studies of Rahwan et al. [35] and ourselves were such that the conclusion of argument B attacks a premise of argument A and, likewise, the conclusion of argument C attacks a premise of argument B. Consider again the car battery example from Section 2. It is not obvious that the attacks of B on A and of C on B are asymmetric: since the conclusions and premises involved in these attacks are contradictory, the attacks might also be regarded as symmetric. This is, for instance, possible in *ASPIC+* [26] in which a so-called 'ordinary' premise can rebut an argument with a 'defeasible top rule'. Moreover, *AF*s generated in *ASPIC+* include the subarguments of all arguments as separate arguments, including arguments corresponding to a premise.

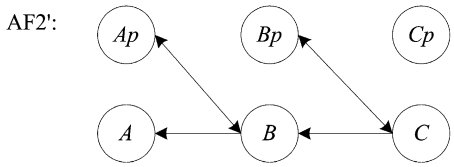

Fig. 4. An alternative interpretation of Rahwan et al.'s examples.

Thus another plausible AF modelling of the car battery example is AF_2' as shown in Fig. 4, where Ap, Bp, and Cp are the subarguments of, respectively, A, B, and C, consisting of their premise. Note that B and Ap attack each other since B undermines Ap (and A) while Ap rebuts B. Likewise for the other attacks. Note also that unlike in $AF2$, in $AF2'$ argument A is not skeptically acceptable. Now it is important to note that a number of participants may have interpreted the examples as in $AF2'$ instead of $AF2$. In an experiment conducted by Cramer and Guillaume [10] this was indeed found to be the case. The participants who did so may have rated the conclusion of A lower in the reinstatement stage since A is only credulously acceptable in that stage.

This is an instance of a more general problem with this kind of empirical research. In experiments like these, a natural-language reasoning example is formalised, then humans are asked to express an opinion of the natural-language version of the example, and then the humans' responses are compared to the outcome yielded by the semantics of the formalised version of the example. If there is a mismatch between the two, then it is tempting to conclude that humans do not reason according to the formal semantics but such a conclusion is premature, since the mismatch may also be caused by the fact that the formalisation does not correspond to what the humans had in mind (this point is also made by Polberg and Hunter [31]). Formalising informal reasoning examples is far from trivial since natural language is ambiguous and the same informal way of reasoning may be formalised in the same formalism in different ways. The danger of such mismatches between a formalised example and how humans interpret its natural-language version increases the more abstract the formalism is. As noted by Prakken and de Winter [34], directly formalising natural-language examples in abstract argumentation formalisms without being guided by a theory of the nature of arguments and their relations may result in ad-hoc modellings (or in the present case in a modelling that is not the only possible one).

This danger may also have materialised in a study of Rosenfeld and Kraus [36], who modelled natural-language examples in a bipolar argumentation framework (an AF with also support relations) and then observed that the participants did not assess arguments according to its semantics, including the reinstatement pattern. This result was cited by [1] as support for gradual argumentation semantics. However, in Rosenfeld and Kraus's examples several attack relations modelled as asymmetric can also be regarded as symmetric. For example, the arguments "We should buy an SUV; it's the right choice for us" and "But we can't afford an SUV, it's too expensive" (where according to Rosenfeld and Kraus the second asymmetrically attacks the first) could by some participants be regarded as two arguments with contradictory conclusions 'we should buy an SUV' and 'we should not buy an SUV'.

A related problem with such empirical reasoning experiments is that it is often hard to make the participants stick to the information that was explicitly given; often they will, either implicitly or explicitly, also take other beliefs and background information into account. Benthem [3] (cited by Rahwan et al. [35] in support of the relevance of empirical research for normative theories) notes that people in such experiments first go through a 'representation' or 'modelling' phase in which they construe the relevant scenario of facts and events, and only then make inferences from the construed scenario. He points to the possibility that experimenters overlook that the participants may have added information to the example

in the representation phase. Other recent empirical studies in computational argumentation have also pointed at the possibly confounding effect of background information [8,11,12,21,31].

In the present study we tried to avoid the unwanted influence of background information as follows. Overall, the arguments that were used were simple sentences and of a neutral subject matter, to avoid unwanted influence of subjective views. Moreover, the levels of confidence in the eighth set (i.e., the one regarding animal rights, which is not a neutral subject matter), do not deviate from the rest in any way. This suggests a good level of impartiality from the participants because the eighth set is deontic and hence formally different from the others, since all the other sets are about truth and this one is about ethical norms. Nevertheless, we cannot exclude the possibility that the results are partly influenced by the *content* of the arguments rather than their *relations*, since content effects are often observed in tasks where they should not (for instance, on the Wason selection task [9]). In order to render such experiments less precarious, future empirical research could try to control for such issues by including manipulation checks, where separate groups of participants evaluate the arguments independently, indicate how they perceive the type and the directionality of the attacks, and so on.

4.4. Cognitive load & order

There are further possible explanations of some of the findings. First, the results of the second group could also be explained by the task of group 2 (i.e., the version of manner of presentation that corresponded to group 2) being more challenging. As mentioned by Cramer and Guillaume [12], a cognitively challenging task might lead to participants choosing a simplifying strategy, in this case, more likely to choose a 'neutral' rating (in this experiment, that would translate to a rating being closer to 4, hence being the lowest rated). The low overall ratings of argument A in group 2, along with the fact that group 2 had the highest dropout rate (26% of the participants of the second group left the survey unfinished, compared to the 21% of the first and then 15% of the third) might be an indication that the manner of presentation in the second group was more challenging to the subjects.

Second, the imperfect recovery from attack could be a result of *order*. For example, the order of presentation may have had an effect on how the participants perceived the directionality of the attacks; it may be that attacks are more often regarded as originating from the last-presented argument. Moreover, it is possible to assume that participants' confidence in A's conclusion does not go back to its original level because the sooner we are introduced to something, the more likely we are to believe it. As observed by Polberg and Hunter [31]: "presenting a new and correct piece of information that a given person was not aware of does not necessarily lead to changing that person's beliefs". Both in our study and in Rahwan et al. [35], the arguments were always presented in the same order. Even in group 2, where all the arguments were presented together, argument A is always first. We cannot, therefore, rule out the possibility that the order of arguments also plays a role in participants' confidence.

4.5. The jump from is to ought

Nevertheless, suppose that future experiments are able to reproduce Rahwan et al.'s findings in examples that unambiguously correspond to Fig. 1 and in which background information has been controlled for. Then there is another hurdle to take before it can be concluded that these results support the void precedence principle as a normative principle of rational argumentation. This hurdle is that it is not immediately obvious how empirical findings on how people *actually* argue can be relevant for a normative theory on how they *should argue*. Given the growing number of empirical studies in computational argumentation [8,10–12,21,29,31,36], this question is important, but it has no simple answers. Another

reason why this question is important is that a 'having fewer attackers is better' principle implies that it is rational for arguers to utter as many counterarguments to an argument as possible, even if these arguments are silly and can be easily refuted. One may wonder whether normative models of argumentation should really encourage arguers to build their arguments on fake news and alternative facts as much as possible.

Rahwan et al. [35] argue that insights from psychological experiments can be relevant to the design of software agents that can argue persuasively with humans. We could think here of IBM's Debater project [37]. They may very well be right in this: persuasiveness is a psychological phenomenon, so psychological experiments can obviously yield relevant insights into the persuasiveness of argumentation patterns. However, in our opinion, formal models like Dung's abstract argumentation theory [13] or more concrete structured accounts like *ASPIC+* [26], Defeasible Logic Programming [18] or Assumption-based Argumentation [39] do not aim to model *persuasiveness* of arguments. Instead, they model the (non-monotonic) *logical* status of arguments as part of a set of arguments and their logical relations of attack and support.

Rahwan et al. also argue that empirical findings on how humans actually argue are relevant for validating formal semantics of argumentation (see also Pfeifer and Tulkki [30], who argue that the process of constructing formal-normative theories can be guided by experimental data in an interactive process). However, they are not explicit on when a formal semantics should be changed because of empirical findings on how humans argue and when humans should change their way of arguing to make it fit the formal semantics. One reason to change the formal semantics might be an assumption that humans by-and-large reason correctly. For example, Pollock [32] argued that the reasoning of humans is guided by internalised rules, while Jackson [24] argued that any descriptive attempt constitutes a "reconstruction of people's own normative ideas". However, it is not obvious why this should always hold. For instance, it has been claimed on the basis of experiments that humans are generally poor at reasoning correctly with and about probabilities (although this claim is not uncontested; see e.g. [14,27]). This claim is generally not regarded as entailing that probability theory is invalid as a normative theory of reasoning with probabilities (here, too, the relevance of background information has been noted; cf. [3]).

One of us has argued in Prakken [33] that there is a weaker sense in which empirical findings on how humans reason can be relevant for normative theories of reasoning. Such normative theories should not only be rationally well-founded but also 'cognitively plausible' in that it is not too difficult for people to adhere to their standards. For this reason, theories of reasoning should be stated in terms that are natural to people, such as argumentation-related concepts. Such cognitively plausible normative theories may still deviate from how people actually reason, as long as they are stated in terms that are natural to people. Therefore, a complete theory of argumentation should combine a descriptive part on what concepts are natural to people with a normative part on how people should argue in terms of these concepts.

Applying this to the present discussion, this means that empirical research can tell us that people tend to assess arguments in gradual terms, so that it is important to develop normative theories of gradual argument evaluation. However, the specific designs of such theories cannot be based on empirical research alone but should also apply philosophical insights. In the case of gradual and ranking-based semantics, these insights must, to the best of our knowledge, still largely be developed. For instance, the only defence of the void precedence principle besides references to empirical findings that we could find is the claim of Amgoud and Ben-Naim [2] that this principle is "natural". We suggest that a philosophical

analysis should aim to clarify what is meant by argument acceptability or argument strength and should take the nature of arguments and their relations into account.

5. Conclusion

In this paper, we returned to the experiments of Rahwan et al. [35] on 'simple reinstatement' patterns in formal argumentation for two reasons. First, we wanted to see whether their results can be replicated. We were able to do so with a considerably larger number of participants, which is a significant result given the current concerns about replicability of results in the social sciences, specifically in social psychology. Second, we wanted to investigate with two variants of Rahwan et al.'s experiments whether empirical findings on how humans evaluate arguments in reinstatement cases can support a special case of the 'fewer attackers is better' principle called 'void precedence', which is incorporated in many current graded notions of argument acceptability. Among other things, we wanted to investigate whether suggestions in the literature that Rahwan et al.'s results can be used to support or criticise principles of argumentation semantics are justified. We can draw the following main conclusions from our investigations.

To start with, our results cast doubt on explanations suggested by Rahwan et al. [35] and Prakken and de Winter [34] in terms of suspension of disbelief. According to these explanations, the imperfect recovery of arguments from attack in reinstatement patterns would be due to the triggering at various moments of awareness or consideration of other counterarguments than those presented in the experiment. In our new experiments we did not find evidence for these explanations.

However, we concluded that this does not imply that the experimental results of Rahwan et al. and the present paper support the 'fewer attackers is better' principle. We first noted that the experiments at best support a special case of this principle, namely, 'an argument is better if it is unattacked than if it is attacked' (void precedence). Next, we concluded that even the special case is not supported since several alternative explanations cannot yet be ruled out, such as that a number of participants may have had different attack relations in mind. More generally, we highlighted the danger that humans involved in reasoning experiments model and/or interpret examples differently than the experimenters. Finally, we argued that even if future experiments extend to the general case and can rule alternative explanations out, still convincing arguments are needed why and how empirical findings on how humans argue can be relevant for normative models of argumentation. We suggested that the importance of such empirical findings does not lie in what they say about the validity of specific reasoning patterns but in what they say about the general concepts that a normative theory should have in order to be applicable by humans. The issue concerning the jump from 'is' to 'ought' is important since the 'having fewer attackers is better' principle implies that it is rational for arguers to utter as many counterarguments to an argument as possible, even if these arguments are silly and can be easily refuted. Should our normative models of argumentation really encourage arguers to build their arguments on fake news and alternative facts as much as possible?

Acknowledgements

This work was partially supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence https://safeandtrustedai.org.

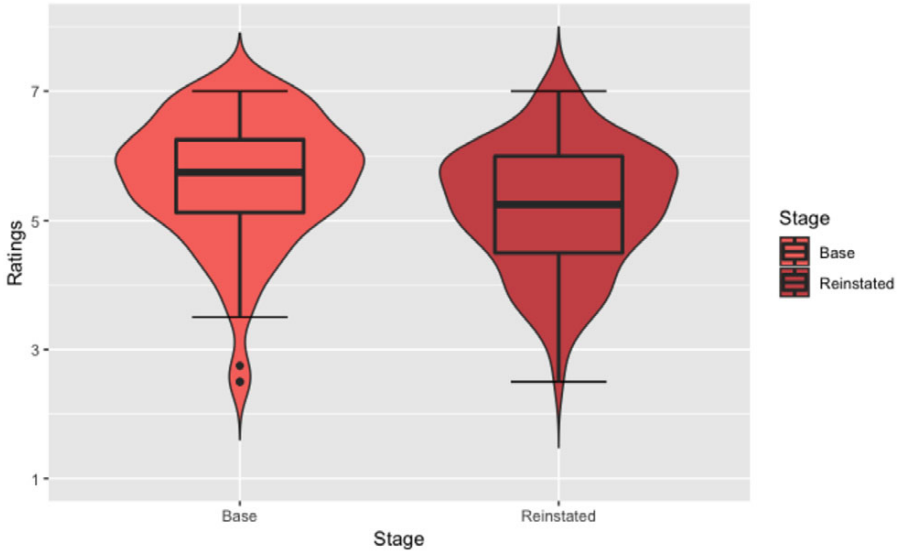

Fig. 5. Experiment 1: ratings of group 1 per stage.

Appendix A. Supplementary graphs

In this section, some additional graphs for the three experiments are presented. They are all a combination of boxplot and violin plot for the ratings of the corresponding condition. In the boxplots, the bold horizontal line segment within the boxes represents the median of the ratings. The violin plot illustrates the distribution of the ratings (the probability density of the data at different values).

In Fig. 5, we see a boxplot and violin plot for the ratings of the first group's participants (y axis) for the two stages; the base stage is depicted on the left, while on the right stands the reinstated stage. The figure shows that the distribution of scores is lower in the reinstated stage compared to the base stage, illustrating the phenomenon of imperfect reinstatement.

In Fig. 6, we see a boxplot and violin plot for groups 1 and 2, where participants' ratings (y axis) are compared; group 1 (reinstated stage) is presented on the left and group 2 is on the right. The figure shows that the distribution of scores is lower in the second group compared to the first but not significantly.

In Fig. 7, we see a boxplot and violin plot for the base stage of groups 1 and 3, where participants' ratings (y axis) are compared; group 1 is depicted on the left and group 3 is on the right. The figure shows that the distribution of scores is lower in the first group compared to the third.

In Fig. 8, we see a boxplot and violin plot for the reinstated stage of groups 1 and 3, where participants' ratings (y axis) are compared; group 1 is depicted on the left and group 3 is on the right. The figure shows that the distribution of scores is lower in the first group compared to the third.

A.1. Follow up comparisons

In Fig. 9, we see a boxplot and violin plot for the ratings within group 3; on the left the base stage is depicted, whereas on the right stands the reinstated. The figure shows that the distribution of scores is lower in the reinstated stage compared to the base stage, illustrating that the phenomenon of imperfect reinstatement also appeared in experiment 3.

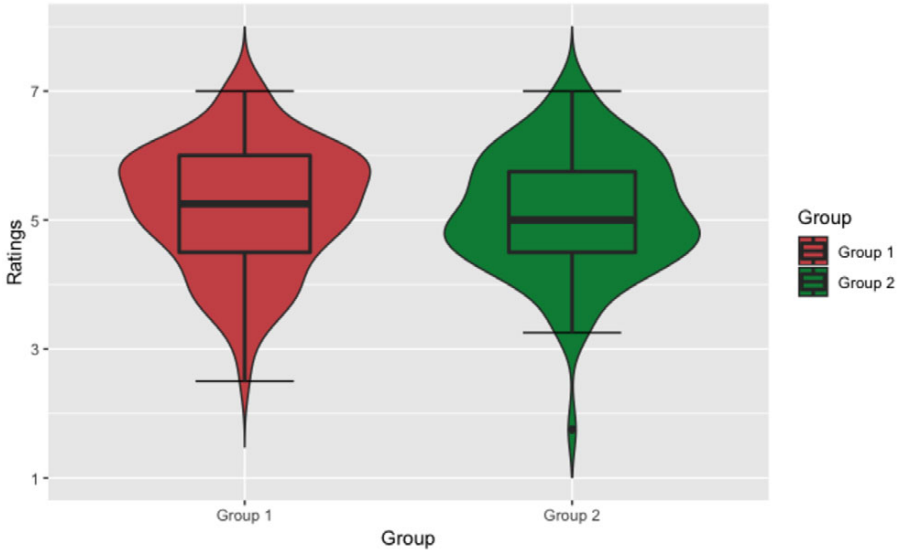

Fig. 6. Experiment 2: ratings of group 1 vs group 2.

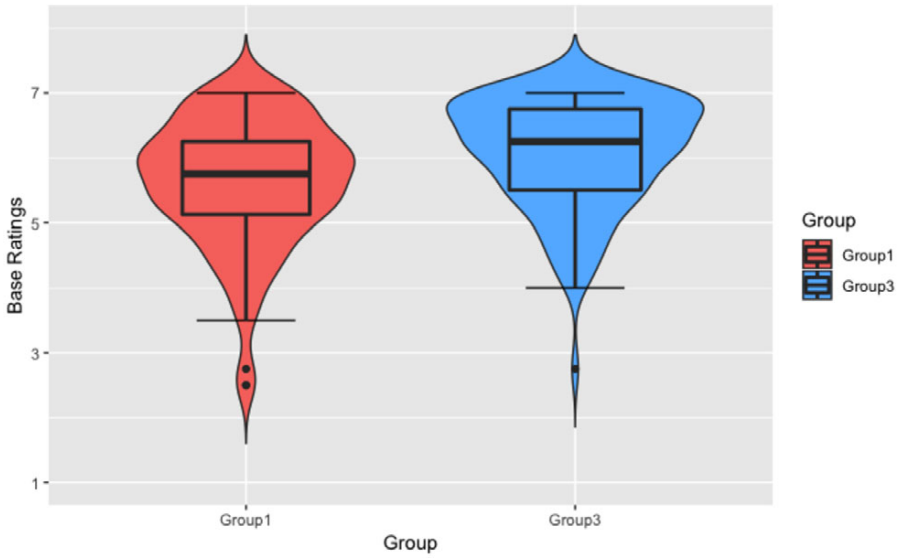

Fig. 7. Experiment 3: base ratings of group 1 vs group 3.

In Fig. 10, we see a boxplot and violin plot for the ratings of all three groups. On the left group 1 (reinstated stage) is depicted, in the middle there is group 2, whereas on the right stands group 3 (reinstated stage). The figure shows that the distribution of scores is the lowest in the second group and the highest in the third.

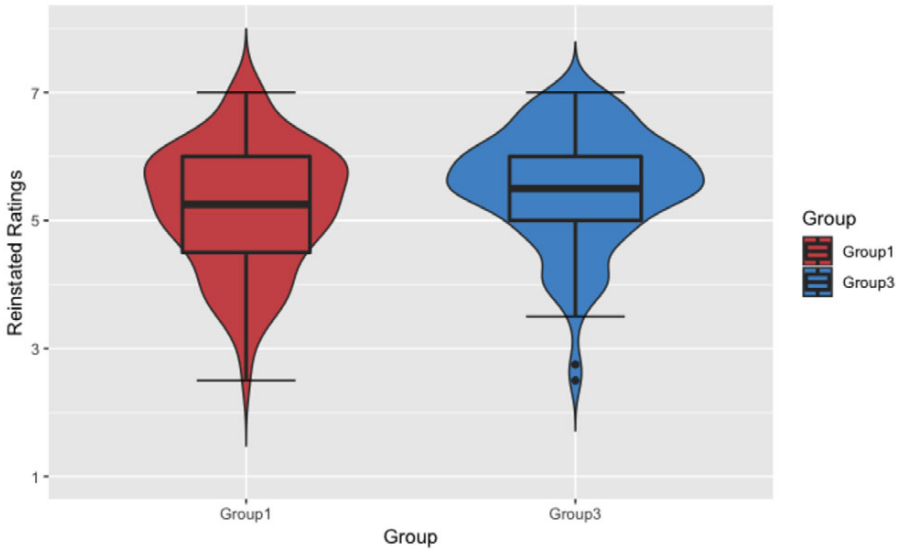

Fig. 8. Experiment 3: reinstated ratings of group 1 vs group 3.

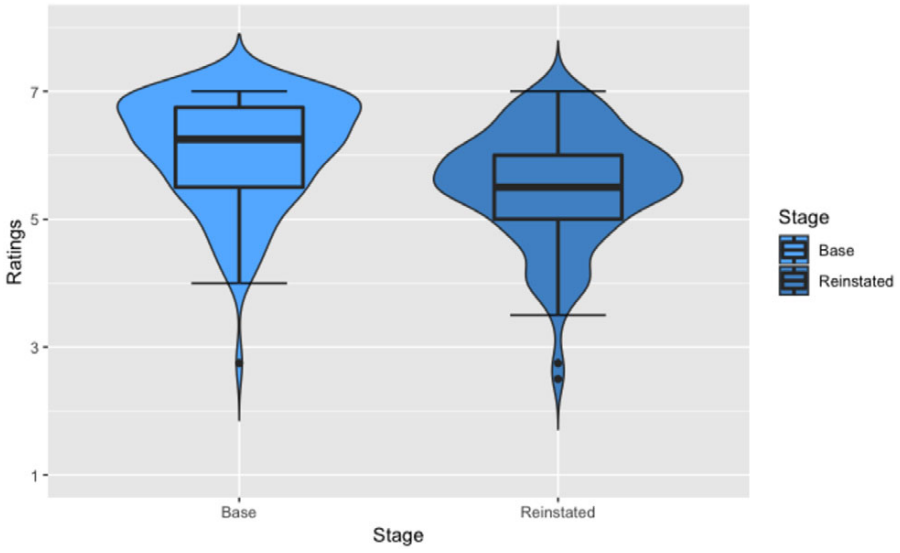

Fig. 9. Ratings of group 3 per stage.

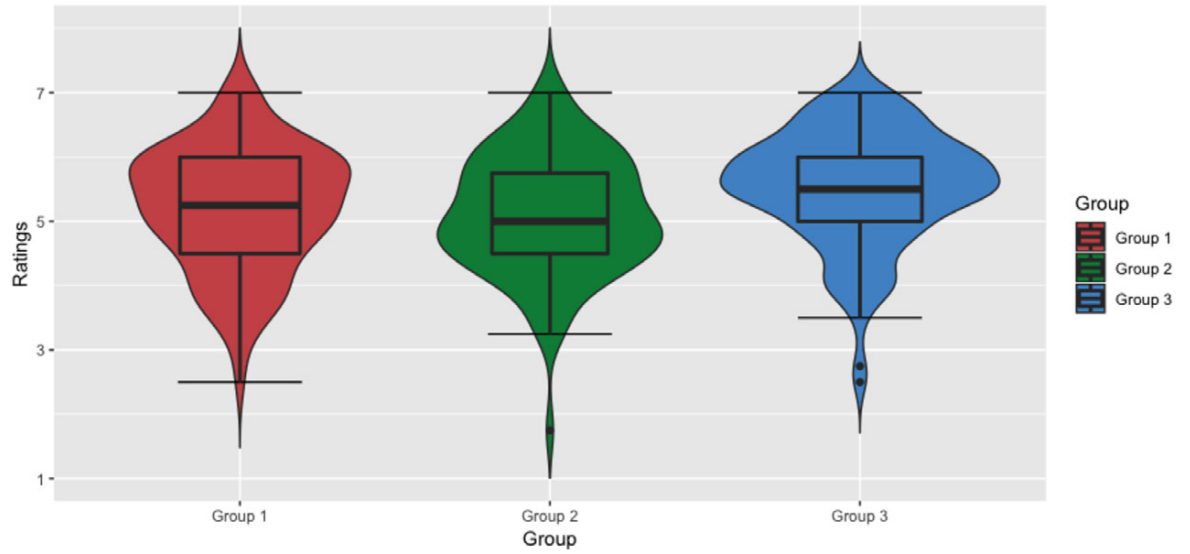

Fig. 10. Ratings of all three groups.

Appendix B. Language questionnaire

The following questions were included in the survey that the participants had to answer prior to participating in the experiment, in order to ensure they possessed the appropriate language level in Englsih.

(1) Are you a native English speaker?
(2) What is your country of residence?
(3) How frequently do you use English during a week (in writing and reading)?
(4) Have you ever taken a formal language test for English?
If yes: Which one and what was the result (what is your level according to it)?

Appendix C. The argument sets

C.1. Original version

Sets 1–6 are taken from Rahwan et al. [35].

Set 1.

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

Set 2.

A: There is no electricity in the house. Therefore, all lights in the house are off.
B: There is a working portable generator in the house. Therefore, there is electricity in the house.
C: The fuel tank of the portable generator is empty. Therefore, the portable generator is not working.

Set 3.

A: Mary does not limit her phone usage. Therefore, Mary has a large phone bill.
B: Mary has a speech disorder. Therefore, Mary limits her phone usage.
C: Mary is a singer. Therefore, Mary does not have a speech disorder.

Set 4.

A: John has no way to know Leila's password. Therefore, Leila's e-mails are secured from John.
B: Leila's secret question is very easy to answer. Therefore, John has a way to know Leila's password.
C: Leila purposely gave a wrong answer to her secret question. Therefore, Leila's secret question is not very easy to answer.

Set 5.

A: Mike's laptop does not have anti-virus software installed. Therefore, Mike's laptop is vulnerable to computer viruses.
B: Nowadays anti-virus software is always available by default on purchase. Therefore, Mike's laptop has anti-virus software.
C: Some laptops are very cheap and have minimal software. Therefore, anti-virus software is not always available by default.

Set 6.

A: Louis applied the brake, and the brake was not faulty. Therefore, the car slowed down.
B: The brake fluid was empty. Therefore, the brake was faulty.
C: The car had just undergone maintenance service. Therefore, the brake fluid was not empty.

Set 7.

A: The power is out, so Claire cannot charge her phone.
B: The TV is playing, so the power is not out.
C: The TV is broken, so the TV is not playing.

Set 8.

A: Animals have the right to be left unharmed, so we should ban animal testing.
B: Animals are very dissimilar to humans, so animals do not have such a right.
C: Animals resemble us anatomically, physiologically, and behaviourally (e.g., recoiling from pain, fearing tormentors), therefore they are not very dissimilar to humans.

C.2. Generalisations of the argument sets

Corresponding generalisations of the argument sets above, used for the third group.

Set 1.

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

Set 2.

When there is no electricity in the house, all lights are off.
If there is a working portable generator in the house, there is electricity in the house.
When the fuel tank of a portable generator is empty, the generator is not working.

Set 3.

When one uses their phone a lot, they have a large phone bill.
When one has a speech disorder, they limit their phone usage.
If someone is a singer, they cannot have a speech disorder.

Set 4.

If someone has no way to know your password, your e-mails are secured from them.
There is a way for someone to know your password if your secret question is very easy.
If one has purposely given a wrong answer to their secret question, that question is not very easy to answer.

Set 5.

If a laptop has no anti-virus software installed, it is vulnerable to computer viruses.
If anti-virus software is always available by default on purchase, all laptops have it.
If there exist some laptops with minimal software, anti-virus software is not always available by default.

Set 6.

When a non-faulty brake is applied, a car slows down.
A brake is faulty if the brake fluid is empty.
When a car has just undergone maintenance service, the brake fluid is not empty.

Set 7.

When the power is out, one cannot charge their phone.
If a TV is playing, the power is not out.
If a TV is broken, it cannot be playing.

Set 8.

If a being has the right to be left unharmed, we should not perform tests on it.
If a being is very dissimilar to humans in order to be able to engage in such a contract, it does not have such a right.
If a being resembles us in various aspects, it is not very dissimilar to humans.

Appendix D. Materials presentation

The presentation of material was the following. The first pages consisted of a brief example of an argument and the instructions about the procedure, where it was explicitly emphasised that participants

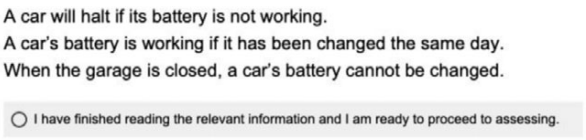

Fig. 11. The relevant theory of set 1 (used only for group 3).

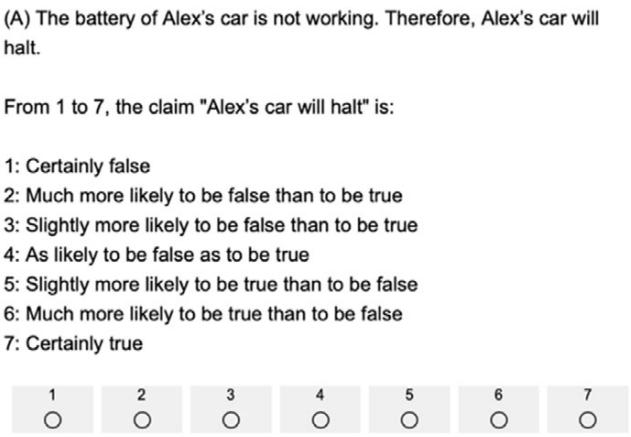

Fig. 12. Base stage of set 1.

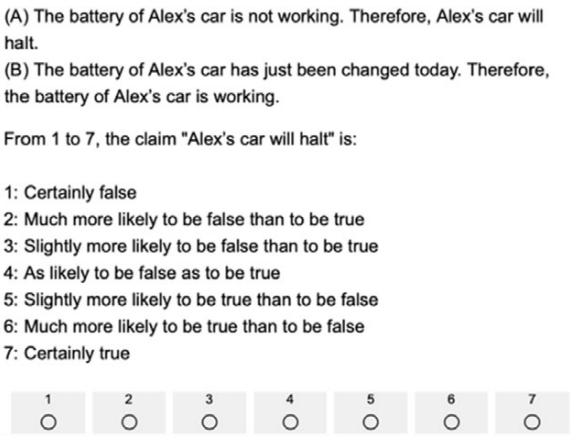

Fig. 13. Defeated stage of set 1.

had to stick solely to the provided information for evaluating the conclusion of argument A. Then, each participant was presented with four (out of the eight) different sets of arguments and asked to evaluate the conclusion of argument A on a 7–point scale, ranging from *Certainly false* to *Certainly true*. In Figs 11, 12, 13, 14, the parts of the survey for each stage of the first argument set are illustrated. Each stage was presented on a different page and no 'go back' option was available. Group 1 was sequentially presented with 12, 13, and 14. Group 2 was only presented with 14. Group 3 was first presented with 11 and then, same as group 1, sequentially with 12, 13, and 14.

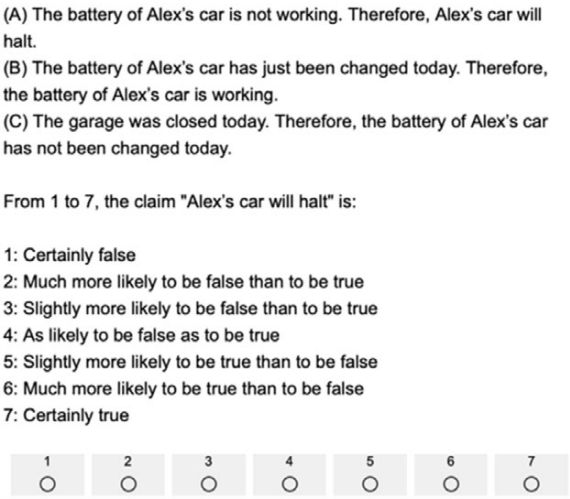

Fig. 14. Reinstated stage of set 1.

References

[1] L. Amgoud, A replication study of semantics in argumentation, in: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019, pp. 6260–6266.
[2] L. Amgoud and J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: *Scalable Uncertainty Management. SUM 2013*, W. Liu, V. Subrahmanian and J. Wijsen, eds, Springer Lecture Notes in Computer Science, Vol. 8078, Springer Verlag, Berlin, 2013, pp. 134–147.
[3] J.V. Benthem, Logic and reasoning: Do the facts matter?, *Studia Logica* **88** (2008), 67–84. doi:10.1007/s11225-008-9101-1.
[4] E. Bonzon, J. Delobelle, S. Konieczny and N. Maudet, A comparative study of ranking-based semantics for abstract argumentation, in: *Proceedings of the 30st AAAI Conference on Artificial Intelligence (AAAI 2016)*, 2016, pp. 914–920.
[5] E. Bonzon, J. Delobelle, S. Konieczny and N. Maudet, A parametrized ranking-based semantics compatible with persuasion principles, *Argument and Computation* **12** (2021), 49–85. doi:10.3233/AAC-200905.
[6] P. Cairns, *Doing Better Statistics in Human–Computer Interaction*, Cambridge University Press, 2019.
[7] M. Caminada, On the issue of reinstatement in argumentation, in: *Logics in Artificial Intelligence. Proceedings of JELIA 2006*, M. Fischer, W. van der Hoek, B. Konev and A. Lisitsa, eds, Springer Lecture Notes in AI, Vol. 4160, Springer Verlag, Berlin, 2006, pp. 111–123.
[8] F. Cerutti, N. Tintarev and N. Oren, Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation, in: *Proceedings of the 21st European Conference on Artificial Intelligence*, 2014, pp. 207–212.
[9] L. Cosmides, The logic of social exchange: Has natural selection shaped how humans reason? Studies with the wason selection task, *Cognition* **31**(3) (1989), 187–276. doi:10.1016/0010-0277(89)90023-1.
[10] M. Cramer and M. Guillaume, Directionality of attacks in natural language argumentation, in: *Proceedings of the Fourth Workshop on Bridging the Gap Between Human and Automated Reasoning*, 2018, pp. 40–46.
[11] M. Cramer and M. Guillaume, Empirical cognitive study on abstract argumentation semantics, in: *Computational Models of Argument*, S. Modgil, K. Budzynska and J. Lawrence, eds, Proceedings of COMMA 2018, IOS Press, Amsterdam, 2018, pp. 413–424.
[12] M. Cramer and M. Guillaume, Empirical study on human evaluation of complex argumentation frameworks, in: *Proceedings of the 16th European Conference on Logics in Artificial Intelligence (JELIA 2019)*, F. Calimeri, N. Leone and M. Manna, eds, Springer Lecture Notes in AI, Vol. 11468, Springer Verlag, Berlin, 2019, pp. 102–115.
[13] P. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n–person games, *Artificial Intelligence* **77** (1995), 321–357. doi:10.1016/0004-3702(94)00041-X.
[14] J.S.B. Evans and D.E. Over, *Rationality and Reasoning*, Psychology Press, 2013.
[15] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, SAGE Publications, 5th edn, 2021, pp. 169–190, Chapter on ANOVA.
[16] A. Field, *Discovering Statistics Using R*, SAGE Publications, 5th edn, 2021, pp. 123–145, Chapter on T-Test.

[17] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, SAGE Publications, 5th edn, 2021, pp. 146–168, Chapter on Welch's Test.
[18] A. Garcia and G. Simari, Defeasible logic programming: An argumentative approach, *Theory and Practice of Logic Programming* **4** (2004), 95–138. doi:10.1017/S1471068403001674.
[19] D. Grossi and S. Modgil, On the graded acceptability of arguments, in: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 868–874.
[20] D. Grossi and S. Modgil, On the graded acceptability of arguments in abstract and instantiated argumentation, *Artificial Intelligence* **275** (2019), 138–173. doi:10.1016/j.artint.2019.05.001.
[21] M. Guillaume, M. Cramer, L. van der Torre and C. Schiltz, Reasoning on conflicting information: An empirical study of formal argumentation, *Plos one* **17**(8) (2022), e0273225.
[22] A. Hunter, S. Polberg and M. Thimm, Epistemic graphs for representing and reasoning with positive and negative influences of arguments, *Artificial Intelligence* **281** (2020), 103236.
[23] A. Hunter and M. Thimm, Probabilistic reasoning with abstract argumentation frameworks, *Journal of Artificial Intelligence Research* **59** (2017), 565–611. doi:10.1613/jair.5393.
[24] S. Jackson, What can argumentative practice tell us about argumentation norms? in: *Norms in Argumentation*, R. Maier, ed., Proceedings of the Conference on Norms, Foris Publication, Dordrecht/Providence RI, 1989, pp. 113–122. doi:10.1515/9783110877175-010.
[25] H. Jakobovits, On the Theory of Argumentation Frameworks, Doctoral dissertation, Free University Brussels, 2000.
[26] S. Modgil and H. Prakken, The ASPIC+ framework for structured argumentation: A tutorial, *Argument and Computation* **5** (2014), 31–62. doi:10.1080/19462166.2013.869766.
[27] M. Oaksford and N. Chater, New paradigms in the psychology of reasoning, *Annual review of psychology* **71** (2020), 305–330. doi:10.1146/annurev-psych-010419-051132.
[28] H. Pashler and E. Wagenmakers, Editors' introduction to the special section on replicability in psychological science: A crisis in confidence?, *Perspectives on Psychological Science* **7** (2012), 528–530. doi:10.1177/1745691612465253.
[29] N. Pfeifer and C. Fermüller, Probabilistic interpretations of argumentative attacks: Logical and experimental results, *Argument and Computation* **14** (2023), 75–107. doi:10.3233/AAC-210016.
[30] N. Pfeifer and L. Tulkki, Conditionals, counterfactuals, and rational reasoning: An experimental study on basic principles, *Minds and Machines* **27** (2017), 119–165. doi:10.1007/s11023-017-9425-6.
[31] 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. doi:10.1016/j.ijar.2017.11.009.
[32] J. Pollock, *Contemporary Theories of Knowledge*, Rowman & Littlefield, Littlefield, NY, 1986.
[33] H. Prakken, On validating theories of abstract argumentation frameworks: The case of bipolar argumentation frameworks, in: *Proceedings of the 20th Workshop on Computational Models of Natural Argument*, CEUR Workshop Proceedings, Vol. 2669, 2020, pp. 21–30.
[34] H. Prakken and M. de Winter, Abstraction in argumentation: Necessary but dangerous, in: *Computational Models of Argument*, S. Modgil, K. Budzynska and J. Lawrence, eds, Proceedings of COMMA 2018, IOS Press, Amsterdam, 2018, pp. 85–96.
[35] I. Rahwan, M. Madakkatel, J.-F. Bonnefon, R. Awan and S. Abdallah, Behavioural experiments for assessing the abstract semantics of reinstatement, *Cognitive Science* **34** (2010), 1483–1502. doi:10.1111/j.1551-6709.2010.01123.x.
[36] A. Rosenfeld and S. Kraus, Providing arguments in discussions based on the prediction of human argumentative behavior, in: *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, 2015, pp. 1320–1327.
[37] N. Slonim, Y. Bilu and C. Alzate, An autonomous debating system, *Nature* **591** (2021), 379–384.
[38] M. Thimm and G. Kern-Isberner, On controversiality of arguments and stratified labelings, in: *Computational Models of Argument*, S. Parsons, N. Oren, C. Reed and F. Cerutti, eds, Proceedings of COMMA 2014, IOS Press, Amsterdam, 2014, pp. 413–420.
[39] F. Toni, A tutorial on assumption-based argumentation, *Argument and Computation* **5** (2014), 89–117. doi:10.1080/19462166.2013.869878.