Behavioural Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement

Iyad Rahwan
[1]Masdar Institute of Science & Technology
[2](Visiting Scholar) Massachusetts Institute of Technology
[3](Fellow) University of Edinburgh

Mohammed Iqbal Madakkatel
British University in Dubai

Jean-François Bonnefon
CNRS and Université de Toulouse

Ruqiyabi Naz Awan
British University in Dubai

Sherief Abdallah
[1]British University in Dubai
[2](Fellow) University of Edinburgh

Argumentation is a very fertile area of research in Artificial Intelligence, and various semantics have been developed to predict when an argument can be accepted, depending on the abstract structure of its defeaters and defenders. When these semantics make conflicting predictions, theoretical arbitration typically relies on ad hoc examples and normative intuition about what prediction ought to be the correct one. We advocate a complementary, descriptive-experimental method, based on the collection of behavioural data about the way human reasoners handle these critical cases. We report two studies applying this method to the case of reinstatement (both in its simple and floating forms). Results speak for the cognitive plausibility of reinstatement, and yet show that it does not yield the full expected recovery of the attacked argument. Furthermore, results show that floating reinstatement yields comparable effects to that of simple reinstatement, thus arguing in favour of preferred argumentation semantics, rather than grounded argumentation semantics. Besides their theoretical value for validating and inspiring argumentation semantics, these results have applied value for developing artificial agents meant to argue with human users.

Keywords: Argumentation; conditionals; nonmonotonic reasoning; defeasible reasoning; logic; psychologism.

Introduction

Understanding human reasoning and decision-making is a key question in cognitive science. There is considerable literature on understanding whether and why people deviate from formal, normative models of deductive reasoning (Bonnefon, 2009; Evans & Over, 2004; Johnson-Laird & Byrne, 2002), decision-theoretic reasoning (Shafir & LeBoeuf, 2002; Tversky & Kahneman, 1981), and defeasible reasoning (Stenning & Lambalgen, 2008). The latter involves a basic form of reasoning in the presence of conflicting information, which can also be referred to as *argumentation* (van Eemeren, Grootendorst, & Henkemans, 1996).

Argumentation has become a very fertile area of research in Artificial Intelligence, as illustrated by recent volumes and journal special issues (Bench-Capon & Dunne, 2007; Besnard & Hunter, 2008; Rahwan & McBurney, 2007; Rahwan & Simari, 2009). A highly influential framework for studying argumentation-based reasoning was introduced by Dung (1995). An *argumentation framework* is simply a pair $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ where $\mathcal{A}$ is a set of arguments and $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ is a defeat relation between arguments. This approach focuses on the defeat relations between arguments, leaving aside their origin or their internal structure. Various semantics have attempted to characterise 'correct' argumentation-based reasoning within such a framework. Given an argumentation framework (that can take the form of a graph), a semantics assigns a status to each argument, that is, it determines whether or not the argument can be accepted.[1]

These semantics typically come from a normative perspective, which relies on intuition and ad hoc hypothetical examples as to what constitutes correct reasoning. We will argue that there are limits to relying solely on this approach, and we will advocate the use of psychological experiments as a methodological tool for informing and validating intuitions about argumentation-based reasoning.

In this article, we apply this experimental method to the problem of reinstatement, both in its simple and floating form. All classical semantics deem simple reinstatement to be acceptable, but different semantics have different takes on

[1] Other semantics (e.g., M. W. A. Caminada, 2006a) introduce a more fine-grained distinction between accepted, rejected, and undecided arguments. A comprehensive review of argumentation semantics is beyond the scope of this article, but excellent reviews can be found elsewhere, for example in Baroni and Giacomin (2007) or Rahwan and Simari (2009, Chapter 2).

the special case of floating reinstatement. We will show that psychological experiments can help to evaluate these various semantics, and can provide unique insights even when all formal semantics are in agreement. Not only these insights can inform current and future semantics, but they are relevant to the design of software agents that can argue persuasively with humans, or provide reliable support to human evaluation of arguments (e.g., on top of argument diagrammating tools).

In the next section, we offer a brief reminder of Dung's abstract theory of argumentation, focusing on our examples of choice, simple and floating reinstatement. Then, we discuss how argumentation semantics are typically evaluated in the Artificial Intelligence literature, and we motivate the need for an empirical perspective. We then report two empirical studies investigating simple and floating reinstatement, respectively.

Abstract Argumentation Frameworks

In this section, we summarise key elements of abstract argumentation frameworks. This section contains technical background only, whose outline is the following. Figure 1 displays the canonical graph of simple reinstatement, whereas Figure 2 displays the canonical graph of floating reinstatement. The main question is, in both cases, whether *A* can be accepted. For simple reinstatement, *A* is accepted by preferred as well as grounded semantics. For floating reinstatement, *A* is not accepted by grounded semantics, but is accepted by preferred semantics. Additionally, preferred semantics also accept *C* and *D* in the (formally defined) 'credulous' sense, but not in the 'sceptical' sense.

We now lay bare the technical background required to arrive at these conclusions. In the following, we adopt the common assumption that argument sets are finite, and we begin with Dung's (1995) abstract definition of an argumentation framework.

Definition 1 (Argumentation framework) *An* argumentation framework *is a pair* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *where* $\mathcal{A}$ *is a set of arguments and* $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ *is a defeat relation. An argument* α defeats *an argument* β *iff* $(\alpha, \beta) \in \rightharpoonup$*, also written* $\alpha \rightharpoonup \beta$.

An argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterise defeat among arguments.

The directed graphs displayed in Figures 1 and 2 will be our running examples all through the article. These two graphs display the canonical forms of simple and floating reinstatement, respectively. As it will appear in the course of this section, the critical issue with these examples is whether argument *A* can be accepted in spite of being defeated by argument *B*.

Example 1 *The graph in Figure 1 (simple reinstatement) consists of three arguments A, B, C, and features two defeat relations:* $B \rightharpoonup A$ *and* $C \rightharpoonup B$*. The graph in Figure 2 (floating reinstatement) consists of four arguments A, B, C, and D, and features five defeat relations:* $B \rightharpoonup A$, $C \rightharpoonup B$, $D \rightharpoonup B$, $C \rightharpoonup D$, *and* $D \rightharpoonup C$.

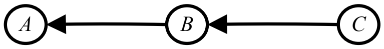

Figure 1. The canonical graph of defeat and simple reinstatement. Argument *A* is defeated by argument *B*, which is in turn defeated by argument *C*.

Note that each node is a complete argument: i.e. a premise as well as a conclusion. The arrows between the nodes represent defeats among arguments. To understand how actual arguments following these graph structures look like, consider the following three arguments that follow the simple reinstatement structure in Figure 1.

(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.

Clearly, argument (B) is an attempt to defeat argument (A) by undermining the latter's main premise –that is, argument (B) concludes that Mary limits her phone usage, negating (A)'s premise that she does not do so. In a similar fashion, argument (C) defeats argument (B) itself by undermining (B)'s premise.

There are many ways to define defeat (Rahwan & Simari, 2009). To simplify the reasoning problem, we opted to go with an explicit and simple notion of defeat: the defeater's conclusion explicitly negates the defeated argument's premise. This, so-called *undercutting* defeat, also insures that the defeats are not symmetric.

The following natural language arguments follow the floating reinstatement structure shown in Figure 2.

(A) Cody does not fly. Therefore, Cody is unable to escape by flying.

(B) Cody is a bird. Therefore, Cody flies.

(C) Cody is a rabbit. Therefore, Cody is not a bird.

(D) Cody is a cat. Therefore, Cody is not a bird.

Note here that (B) defeats (A) as above. Both (C) and (D) defeat (B) by undercutting its premise that Cody is a bird. However, (C) and (D) mutually defeat each other, since their conclusions are contradictory (so-called *rebutting* defeat).

We now need to define the two fundamental notions of conflict-freedom and defence. First, we introduce the notations S^+ and α^-. For a given set S of arguments, S^+ is the set of arguments that are defeated by the arguments in S. Formally, $S^+ = \{\beta \in \mathcal{A} \mid \alpha \rightharpoonup \beta \text{ for } \alpha \in S\}$. Conversely, for a given argument α, the set α^- is the set of all arguments that defeat α. Formally, $\alpha^- = \{\beta \in \mathcal{A} \mid \beta \rightharpoonup \alpha\}$.

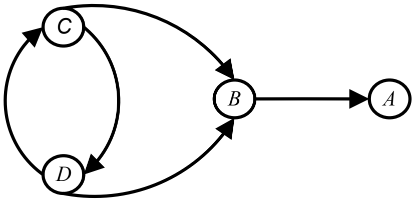

Figure 2. The canonical graph of defeat and floating reinstatement. Argument A is defeated by B, which is itself defeated by C as well as D, although C and D are mutual defeaters.

Definition 2 (Conflict-freedom) *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework and let $S \subseteq \mathcal{A}$. S is* conflict-free *iff $S \cap S^+ = \emptyset$.*

In other terms, a set of arguments is *conflict free* if and only if no argument in that set defeats another.

Definition 3 (Defence) *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, let $S \subseteq \mathcal{A}$, and let $\alpha \in \mathcal{A}$. S* defends *α if and only if $\alpha^- \subseteq S^+$. We also say that argument α is* acceptable *with respect to S.*

In other terms, a set of arguments *defends* a given argument if and only if it defeats all its defeaters.

Example 2 *In the graph displayed in Figure 1, the set $\{A, C\}$ is conflict free, but the set $\{A, B\}$ is not, and neither is the set $\{B, C\}$. Because the set $\{C\}$ defeats all the defeaters of A, we can say that the set $\{C\}$ defends argument A. In the graph displayed in Figure 2, the only conflict-free sets (apart from trivial ones containing single arguments) are $\{A, C\}$ and $\{A, D\}$. Either one of the sets $\{C\}$, $\{D\}$, or $\{C, D\}$, defends A against all its defeaters.*

We now define the *characteristic function* of an argumentation framework.

Definition 4 (Characteristic function) *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. The* characteristic function *of AF is $\mathcal{F}_{AF}$: $2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ such that, given $S \subseteq \mathcal{A}$, we have $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} \mid S \text{ defends } \alpha\}$.*

Applied to an argument set S, the characteristic function returns the set of all arguments defended by S. Because we are only dealing in this article with one argumentation framework at a time, we will use the notation $\mathcal{F}$ instead of $\mathcal{F}_{AF}$.

We now turn to various so-called *extensions* that can characterise the collective acceptability of a set of arguments. Essentially, these extensions provide different possible ways to group self-defending arguments together. These extensions will be used subsequently to define the argument evaluation criteria that we study empirically in this paper.

Definition 5 (Complete/grounded/preferred extensions) *Let S be a conflict-free set of arguments in framework $\langle \mathcal{A}, \rightharpoonup \rangle$.*

- *S is a* complete *extension iff $S = \mathcal{F}(S)$.*
- *S is a* grounded *extension iff it is the minimal complete extension with respect to set inclusion.*
- *S is a* preferred *extension iff it is a maximal complete extension with respect to set inclusion.*

S is a complete extension if and only if *all* arguments defended by S are also in S (that is, if S is a fixed point of the operator $\mathcal{F}$). There may be more than one complete extension, each corresponding to a particular consistent and self-defending viewpoint.

Example 3 *In the graph displayed in Figure 1, the set $\{C\}$ is not a complete extension, because it defends A without including it. The set $\{B\}$ is not a complete extension because it includes B without defending it against C –see Figure 3. The only complete extension is $\{A, C\}$. The graph displayed in Figure 2 has two complete extensions, $\{A, C\}$ and $\{A, D\}$ –see Figure 4.*

A grounded extension contains all the arguments in the graph that are not defeated, as well as all the arguments which are defended directly or indirectly by non-defeated arguments. This can be seen as a non-committal view (characterised by the *least* fixed point of $\mathcal{F}$). As such, there always exists a unique grounded extension.

More intuitively, computing arguments in the grounded extension can be seen as a process of labelling nodes of the graph. First, nodes that have no defeaters are labelled 'undefeated' (and included in the extension) and the nodes attacked by them are labelled 'defeated' (and discarded of the extension). Then, all labelled arguments are suppressed and the process is repeated on the resulting sub-graph, and so forth. If no initial, undefeated node can be found for some iteration, all unlabelled nodes are labelled as 'defeated' and the process is terminated.

Example 4 *The graph displayed in Figure 1 has only one complete extension, $\{A, C\}$, which is also its grounded extension. The graph displayed in Figure 2 has two complete extensions $\{A, C\}$ and $\{A, D\}$, but none of this is the grounded extension, because there is no node in the graph that is initially undefeated. In that case, the grounded extension is the empty set.*

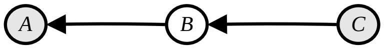

Figure 3. Single (complete, grounded, and preferred) extension in simple reinstatement. Accepted arguments are shaded.

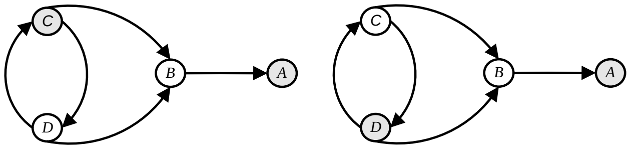

Figure 4. The two (complete, preferred) extensions in floating reinstatement. Accepted arguments are shaded.

A preferred extension is a bolder, more committed position that cannot be extended (by accepting more arguments) without causing inconsistency. Thus a preferred extension can be thought of as a maximal consistent set of hypotheses. There may be multiple preferred extensions, and the grounded extension is included in all of them.

Example 5 *The graph displayed in Figure 1 has only one complete extension, $\{A, C\}$, which is also a preferred extension. The graph displayed in Figure 2 has two complete extensions $\{A, C\}$ and $\{A, D\}$, and both qualify as preferred extensions.*

Now that we have defined various semantics that identify the extensions of an argument graph, we can at last define the status of an individual argument within the graph, that is, we can define criteria for accepting or not each individual argument. The main question in this paper is whether people evaluate a reinstated argument sceptically or credulously in accordance with the definition below.

Definition 6 (Argument status) *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, and $\mathcal{E}_1, \ldots, \mathcal{E}_n$ its extensions under a given semantics. Let $\alpha \in \mathcal{A}$.*

- *α is accepted in the sceptical sense iff $\alpha \in \mathcal{E}_i$, $\forall \mathcal{E}_i$ with $i = 1, \ldots, n$.*
- *α is accepted in the credulous sense iff $\exists \mathcal{E}_i$ such that $\alpha \in \mathcal{E}_i$.*
- *α is rejected iff $\nexists \mathcal{E}_i$ such that $\alpha \in \mathcal{E}_i$.*

Under the grounded semantics, any argument that belongs to the unique grounded extension is accepted both in the credulous and the sceptical sense, and any argument that does not belong to the unique grounded extension is rejected. Under the preferred semantics, an argument is sceptically accepted if it belongs to all preferred extensions; but it can also be credulously accepted if it belongs to at least one preferred extension. If an argument is neither sceptically nor credulously accepted, it is rejected.

Example 6 *The graph displayed in Figure 1 has only one complete extension, $\{A, C\}$, which is grounded as well as preferred. As a consequence, arguments A and C are accepted by grounded as well as preferred semantics, both in the credulous and sceptical sense. The graph displayed in Figure 2 has an empty grounded extension, which means that no argument should be accepted under a grounded semantics. Under a preferred semantics, though, two extensions are identified, $\{A, C\}$ and $\{A, D\}$. From these extensions, only A can be accepted in a sceptical sense, but A, C, and D can all be accepted in a credulous sense.*

What Validates a Semantics?

As established in the previous section, different semantics can have different takes on which arguments can be accepted within a given argumentation framework. The question then arises of evaluating the different claims made by different semantics as to what constitutes an acceptable argument. In this section, we discuss this issue in the broader context of the general sources of inspiration and validation found for these semantics in the formal argumentation literature. We discuss in turn the example-based approach, the principle-based approach, and lastly the experiment-based approach that we suggest needs more attention from the Artificial Intelligence community.

The example-based approach

Most semantics for argumentation-based reasoning in Artificial Intelligence are based on intuition as to what constitutes correct reasoning. A typical research article presents scenarios that can be hypothetical or real (e.g., from the legal domain), and that correspond to one or several argument structures (e.g., floating reinstatement). The proposed semantics is then shown to draw intuitively satisfying conclusions. The difficulty, then, is that one is often able to construct other examples with the same logical structure, in which the proposed semantics draws counter-intuitive conclusions. For example, Horty (2002) famously devoted a whole paper to demonstrate counter-intuitive results with floating conclusions in default reasoning (see also Bonnefon, 2004).

Such counter-intuitive results motivate work on new semantic criteria to capture the novel examples, and the process repeats, examples always being the main tool for comparing semantics with one another. This *example-based approach* (to borrow a term from Baroni & Giacomin, 2007) was, for example, the inspiration for the CF2 semantics (Baroni, Giacomin, & Guida, 2005), dealing with odd-length cycles examples that were problematic for preferred semantics; or for the semi-stable semantics (M. W. A. Caminada, 2006b), dealing with cases in which no stable extension exists, and are shown to have guaranteed existence.

Baroni and Giacomin (2007) made a compelling case for the limitations of the example-based approach, noting in particular that even in relatively simple examples, there might not be a consensual intuition on what should be the correct conclusion. In parallel, Prakken (2002) observed that intuitions about given examples were helpful for generating new investigations, but less helpful as critical tests between different semantics. This recognised difficulty in relying on intuition alone as the benchmark for designing and evaluating argumentation semantics motivated a number of authors to advocate a more systematic approach to which we now turn.

The principle-based approach

To overcome the limitations of the example-based approach, a number of authors recently advocated a more systematic, axiomatic, *principle-based* approach (e.g. Baroni & Giacomin, 2007; M. Caminada & Amgoud, 2007). In this approach, alternative semantics are evaluated by analysing whether they satisfy certain principles, or quality postulates.

Baroni and Giacomin (2007) offered for example the *reinstatement criterion*, according to which an argument must be included in any extension that reinstates it, and *directionality criterion* which requires that an argument's status should only be affected by the status of its defeaters. The Baroni and Giacomin (2007) article offers many other interesting criteria to provide a comprehensive and systematic comparison between abstract argumentation semantics. In parallel, M. W. A. Caminada (2006a) provided postulates for the notion of reinstatement, in order to characterise the labelling of arguments in an argument graph (in, out, and undecided). One postulate states that an argument must be 'in' if and only if all of its defeaters are 'out.' Another postulate states that an argument must be 'out' if and only if at least one of its defeaters is 'in.' This enabled Caminada to characterise different semantics by examining the kinds of labellings they allowed.

The principle-based approach provides a significant improvement over the basic example-based approach, since it enables claims that transcend individual examples and characterise semantics more generally. The source of the general postulates, however, is still the researcher's intuition as to what correct reasoning ought to be. In sum, most of the extent validation of various argumentation semantics, example-based or principle-based, relies on normative claims based on intuition. We now suggest that this normative-intuitive perspective could be adequately complemented with descriptive, *experimental* evidence about how people actually reason from conflicting arguments.

The experiment-based approach

There is a growing concern within the Artificial Intelligence community that logicians and computer scientists ought to give serious attention to cognitive plausibility when assessing formal models of reasoning, argumentation and decision-making. For example, Benthem (2008) strongly supports the rise of a *new psychologism* in logic at large, arguing that although logicians and computer scientists have tended to go by intuition and anecdotal evidence, formal theories can be modified under pressure from evidence obtained though careful experimental design. In the context of epistemic logic, Pietarinen (2003) argues for the important role of empirical findings from cognitive science in revising our logical conceptions of knowledge and belief, commenting that the interplay between logic and cognition is likely to reach increasingly wider and become increasingly prominent.

Pelletier and Elio (1997, 2005) also argued extensively for the importance of experimental data when formalizing default and inheritance reasoning, arguing that default reasoning is particularly psychologistic in that it is *defined* by what people do. Their own results have been complemented by a dynamic experimental literature consisting of controlled tests of human default reasoning (e.g., Benferhat, Bonnefon, & da Silva Neves, 2005; Bonnefon, Da Silva Neves, Dubois, & Prade, 2008; Da Silva Neves, Bonnefon, & Raufaste, 2002; Ford, 2004; Ford & Billington, 2000; Pfeifer & Kleiter, 2005, 2009).

Finally, and in close relation to the problems of simple and floating reinstatement that we have introduced in the previous section, Horty (2002) implicitly appealed to descriptive validation when highlighting the issues that floating conclusions raise for sceptical semantics:[2]

> There is a vivid practical difference between the two skeptical alternatives. [...] Which alternative is correct? I have not done a formal survey, but most of the people to whom I have presented this example are suspicious of the floating conclusion (p.64).

We believe that the field of computational argumentation can indeed benefit from the same kind of formal surveys that have been conducted in the field of default reasoning, and that have been generally called for in Artificial Intelligence. To our knowledge, only very few articles have explicitly sought to inform formal models of argumentation with experimental evidence, and these experimental data have only been collected in relation to the specific issue of argumentation-based decision making (e.g., Amgoud, Bonnefon, & Prade, 2005; Bonnefon, Dubois, Fargier, & Leblois, 2008; Dubois, Fargier, & Bonnefon, 2008). What we offer in this article is an experimental investigation of the basic issue of how people reason from the critical argument structures corresponding to simple and floating reinstatement, and whether one of the current available semantics can capture their reasoning.

Study 1: Simple Reinstatement

Study 1 investigates the the basic structure of argument reinstatement. Abstractly, this structure is defined in the following argumentation framework (as displayed in Figure 1): $AF = \langle\{A,B,C\}, B \rightharpoonup A, C \rightharpoonup B\rangle$, in which argument A is attacked by argument B but reinstated by argument C.

Study 1 seeks to answer the following questions: Does the confidence in the conclusion of A decrease when A is defeated by B? Does this confidence then increase when C is introduced alongside A and B? If so, does confidence return to its initial level, that of when A was presented alone?

Method

Twenty participants were randomly approached in offices, shopping malls, and open spaces in Dubai, to take part in Study 1. Participants read an introduction to the task, informing them that the purpose of the experiment was to collect information about how people thought, that the task included no trick question, and that they simply had to mark the answer that they felt correct. Participants were asked about their proficiency with the English language, in order to make sure that it was above a reasonable level. Participants evaluated their proficiency by choosing one of nine terms ranging from Expert to Very Limited. They then solved 18 problems each, following a 3-level, 6-measure within-participant design.

The 3-level independent variable was the *Pattern* of the problem (Base, Defeated, Reinstated). In the Base pattern, participants were only presented with argument A; in the Defeated pattern, participants were presented with arguments

[2] Working within the scope of default logic, Horty gave a specific example to highlight counter-intuitive results in classical reasoning with a floating conclusion supported by two mutually conflicting pieces of evidence.

A and *B*; finally, in the Reinstated pattern, participants were presented with all three arguments *A*, *B*, and *C*.

Participants saw six different versions of each pattern, which used six different sets of contents for the arguments *A*, *B*, and *C* (see Appendix A for a list of all contents). More specifically, half the participants solved the Base (argument *A*), Defeated (arguments *A* and *B*), and Reinstated (arguments *A*, *B* and *C*) problems using the first set of contents, then the Base, Defeated, and Reinstated problems using the second set of contents, and up to the Base, Defeated, and Reinstated problems using the sixth set of contents. The other half did the same, but they started with the sixth set of contents and worked their way down to the first.

Participants had to answer every problem, in the order they appeared in the questionnaire, without peeking at the next problem in the questionnaire. For each problem, participants had to assess the conclusion of argument *A*, using a 7-point scale anchored at *certainly false* and *certainly true*. The scale and the phrasing of the question were similar to that used in Politzer and Bonnefon (2006).[3]

Manipulation Check

An independent sample of 18 participants was recruited to take part in the manipulation check of Study 1. The purpose of the manipulation check was to make sure that the *C* arguments did a good job at defeating the *B* arguments. Without this precaution, we would not be able to interpret the potential effect of *C* arguments on *A* arguments in the main experiment. Participants in the manipulation check solved 12 problems, according to a 2-level, 6-measure designs. For each of the six argument sets, participants assessed their confidence (on a 7-point scale similar to that used in the main study) in the conclusion of *B* when *B* was presented alone, and their confidence in the conclusion of *B* when *B* was presented together with *C*.

Results

Averaging across the 6 contents and 20 participants, the base confidence in the conclusion (when argument *A* is presented alone) was 5.9 (SD = 0.8) whereas confidence in the defeated conclusion (when argument *A* is attacked by argument *B*) was 4.0 (SD = 1.4). Confidence in the reinstated conclusion (when argument *A* is attacked by argument *B* but reinstated by argument *C*) went back up to 5.2 (SD = 1.0).

Confidence in the conclusion was entered as the dependent variable in a repeated-measure analysis of variance, with pattern as a 3-level predictor (Base, Defeated, Reinstated) and 6 measures corresponding to the 6 contents. The multivariate test detects a significant effect of Pattern, $F(2, 18) = 14.1$, $p < .001$, $\eta_p^2 = .61$. This overall effect reflects both an effect of defeat and an effect of reinstatement. As shown by a contrast analysis, ratings in the Base condition were significantly higher than ratings in the Defeated condition, $F(1, 19) = 26.8$, $p < .001$, $\eta_p^2 = .59$; and ratings in the Defeated condition were themselves significantly lower than ratings in the Reinstated condition, $F(1, 19) = 9.9$, $p = .005$, $\eta_p^2 = .34$. Although reinstatement increased the acceptability of a conclusion, the recovery was not perfect. Indeed, the ratings in the Reinstated condition were still significantly lower than the ratings in the Base condition, $F(1, 19) = 9.1$, $p = .007$, $\eta_p^2 = .32$.

The reliable effect of reinstatement must be related to the success of the reinstating manipulation, as shown by the results of the manipulation check. Averaging across the 6 contents, the base confidence in the conclusion of defeaters was 5.1 (SD = 0.8) whereas it was 4.1 (SD = 0.7) for attacked defeaters. A repeated-measure analysis of variance, with pattern as 2-level predictor, and 6 measures corresponding to 6 contents, detected a significant effect of pattern $F(6, 12) = 3.8$, $p = .02$, $\eta_p^2 = .66$.

Results thus support the notions of defeat and reinstatement. That is, confidence in the conclusion of a defeated argument significantly decreases, but it increases when the defeater is itself attacked by a reinstating argument. Results also suggest, however, that a reinstated argument does not fully recover from its defeat, as confidence in its conclusion remains significantly lower than what it was when the argument was presented in isolation. We defer the discussions of these results until after we report the results of Study 2, which extends Study 1 by considering the more complex case of floating reinstatement.

Study 2: Floating Reinstatement

Study 2 offers an experimental comparison of the simple reinstatement structure to the more complex structure known as floating reinstatement, graphically displayed in Figure 2.

In addition to replicating the findings of Study 1, Study 2 seeks to answer the following questions: Does floating reinstatement restore the confidence in the conclusion of argument *A*, and does it do so to the same extent as simple reinstatement? (A 'yes' to both questions would go against the predictions of grounded semantics.) If so, does the effectiveness of floating reinstatement require that participants manifest a preference for either *C* over *D* or *D* over *C*? (A 'yes' would provide support to the predictions of credulous preferred semantics, a 'no' would provide support to the predictions of sceptical preferred semantics.)

Method

Fourty-seven participants were randomly approached in the same circumstances and following the same protocol as in Study 1. They were randomly assigned to two experimental groups corresponding to simple and floating reinstatement, respectively, then solved 12 problems, following a 3-level, 4-measure within-participant design.

The 3-level independent variable was the *Pattern* of the problem (Base, Defeated, Reinstated). In the Base pattern,

[3] The question always refers to the conclusion of argument *A*. For example, for Argument Set 1 in the appendix, the question would be worded "Alex's car will halt is (1) certainly false; (2) much more false than true; (3) slightly more false than true; (4) as false as true; . . . (7) certainly true." Participants responded by checking the corresponding numeral on a graphically depicted scale.

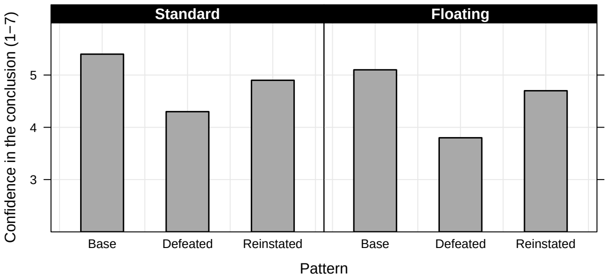

Figure 5. Reinstatement is as effective in its floating form as in its simple form. Confidence in the conclusion of an argument decreases when the argument is defeated, and is then imperfectly restored when its defeater is itself defeated, whether by a single argument (simple reinstatement) or by two mutually defeating arguments (floating reinstatement).

participants were only presented with argument A; in the Defeated pattern, participants were presented with arguments A and B; finally, in the Reinstated pattern, participants were presented with the three arguments A, B, and C (in the simple reinstatement group) or with the four arguments A, B, C, and D (in the floating reinstatement group).

The procedure used in Study 2 was the same as that used in Study 1, but the contents of arguments A, B, C, and D were taken from four different argument sets than in Study 1 (see Appendix B). In addition to the questions used in Study 1, participants rated their understanding of each problem (‘How clearly did you understand the problem?’) on a 7-point scale anchored at *Not at all* and *Completely*. Lastly, participants in the floating reinstatement group answered the following question about the four reinstated problems: Do you think that (i) C is a better argument than D, (ii) D is a better argument than C, or (iii) C and D are about equally good?

Results

Figure 5 displays the average confidence in the conclusion of A, as a function of Pattern and Type of reinstatement, averaged across the contents and participants. The visual inspection of Figure 5 already suggests that the results are very similar for the two groups. This preliminary intuition was confirmed by the results of a mixed-design analysis of variance, using the confidence in the conclusion as a dependent variable, pattern as a 3-level within-subject predictor (Base, Defeated, Reinstated), the type of reinstatement as a 2-level between-group variable (Simple, Floating), and four measures corresponding to the four linguistic contents.

The multivariate test detected a significant effect of Pattern, $F(8, 38) = 6.1$, $p < .001$, $\eta_p^2 = .56$. It did not, however, detect a significant main effect of Type of reinstatement $F(4, 42) < 1$, $p = .79$, $\eta_p^2 = .04$, nor a significant interaction between Pattern and Type, $F(8, 38) = 1.2$, $p = .32$, $\eta_p^2 = .20$.

As in Study 1, the overall effect of Pattern reflected a successful defeat followed by a successful reinstatement. As shown by contrast analysis, confidence ratings in the Defeated condition were significantly lower than ratings in the Base condition, $F(1, 45) = 34.9$, $p < .001$, $\eta_p^2 = .44$, and this difference was not moderated by the Type of reinstatement (there is indeed no reason that it should be), $F(1, 45) < 1$, $p = .67$, $\eta_p^2 < .01$. The confidence ratings in the Reinstated condition were significantly greater than in the Defeated condition, $F(1, 45) = 13.7$, $p < .001$, $\eta_p^2 = .23$, and this difference (more interestingly this time) was not moderated by the Type of reinstatement, $F(1, 45) < 1$, $p = .60$, $\eta_p^2 < .01$. Just as in Study 1, reinstatement is not perfect, as ratings in the Reinstated condition remain significantly lower than in the Base condition, $F(1, 45) = 9.0$, $p < .01$, $\eta_p^2 = .17$. Again, there is no evidence whatsoever of a moderation by Type of reinstatement, $F(1, 45) < 1$, $p = .92$, $\eta_p^2 < .01$.

So far, results suggest that floating reinstatement has an effect that is identical to classic reinstatement. We further note that although subjects found the floating reinstatement problems slightly harder to understand than the simple reinstatement problems, this difference appeared to play no role in the ratings they gave for their confidence in the conclusion. The average understanding rating was 4.6 (SD = 1.1) for simple reinstatement problems, compared to 4.0 (SD = 0.9) for floating reinstatement problems, $t(45) = 2.0$, $p = .05$. However, a regression analysis seeking to predict acceptance of reinstated arguments on the basis of problem understanding, Type of reinstatement (dummy coded, 1 for floating), and the interaction term between these two predictors, failed to find any significant effect. The interaction term in particular achieved a standardized β of .19, non-reliably different from zero, $t = 0.32$, $p = .75$.

The effectiveness of floating reinstatement does not appear to result from the subjects manifesting a preference for one of the mutually defeated arguments. We conducted four repeated-measure analyses of variance, one for each argument set, with conclusion acceptance as a dependent variable, pattern as a 2-level predictor (Defeated, Reinstated), and preference as a dummy coded between-group variable (0 for subjects who said the two mutually defeating arguments were equally good, 1 otherwise). The interaction term between pattern and preference did not achieve statistical significance in any of the four analyses, all Fs in the $0.5 - 1.5$ range, all ps in the $.23 - .48$ range.

General Discussion

Following the introduction of Dung’s (1995) influential *abstract argumentation frameworks,* formal argumentation has become a fertile area of research in Artificial Intelligence. An argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterise defeat among arguments. Within this framework, various semantics (e.g., preferred vs. grounded) have been offered that seek to establish whether or not each argument in the graph can be accepted. In some cases (such as simple reinstatement), preferred and grounded semantics are in agreement; but in other cases (such as floating reinstatement), the two semantics have different takes on what constitutes an acceptable argument.

When there is a conflict between the predictions of two semantics, the standard practice in Artificial Intelligence is to rely on intuition to elect one of these predictions as the normatively correct one. Although this normative-intuitive approach has its uses, we argued that it might be adequately complemented with the kind of descriptive-experimental approach that has already been used in some domains of Artificial Intelligence (e.g., default and inheritance reasoning), and that has been called for by various voices within the formal community. This descriptive experimental approach consists of using the methods of experimental psychology to run controlled studies of argument-based reasoning; and to confront the results of these studies with the predictions made by formal semantics.

In this article, we applied this approach to simple as well as floating reinstatement. Study 1 addressed the basic situation of simple reinstatement, across a varied set of linguistic contents. Participants reasoned in a way that reflected the formal notions of defeat and reinstatement: Their confidence in an argument *A* decreased when it was attacked by an argument *B*, but bounced back up when *B* itself was attacked by a third argument *C*. These findings are in agreement with grounded as well as preferred semantics (and others). What neither semantics could predict, though, is the finding (replicated in Study 2) that the recovery of argument *A* was not complete when reinstated by argument *C*: Confidence in *A* in presence of *B* and *C* did not raise back to its former level, when *A* was presented alone.

This is not a trivial observation. Indeed, every possibility seemed plausible a priori. We could expect, as formal semantics would have it, that *A* would fully regain its former status. We could also imagine that the confidence in *A* in presence of *B* and *C* would surpass the confidence in *A* when presented alone: Indeed, confidence in *A* might be boosted by seeing a potential objection to *A* being ruled out. But what happened was exactly the contrary. Seeing one objection to *A*, even when it was ruled out, decreased the confidence in *A*, possibly because the evocation of one objection prompted participants to consider other possible objections that were not explicitly ruled out in the problem.

There is indeed some sort of *suspension of disbelief* involved in reasoning experiments using natural language materials (see Evans & Over, 2004, Chapter 6, for a review of how to increase or decrease this suspension of disbelief by means of experimental instructions). Participants 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 and some participants start to let their own private beliefs leak into the way they reason from the experimental materials. The fact that simple reinstatement works, though, even if not perfectly, is good news to current semantics, and a warning for future semantics not to dispense with simple reinstatement.

Turning now to floating reinstatement, our results suggest that, empirically speaking, floating reinstatement works exactly as well as simple reinstatement. Participants' confidence in an argument *A* decreased when it was attacked by an argument *B*, but bounced back up when *B* itself was attacked by two mutually defeating arguments *C* and *D*. These results clearly speak in favour of preferred semantics. Results also suggest that the sceptical version of preferred semantics might be more cognitively plausible than the credulous version, since the effect of floating reinstatement was not dependent on participants showing a preference for one of the two mutually defeating arguments. This question is not yet settled, though, since the data do not make it clear whether participants would be willing to commit to accepting one of the mutually defeating arguments *C* and *D*. Hence, this aspect of the results requires further investigation.

Besides their theoretical value, our results also have applied value for developing agents that are meant to argue with human users. We already know that artificial agents can achieve better negotiation results with human users when they do not play normative equilibrium strategies, but rather adopt boundedly rational strategies inspired from human behavioural data (Gal & Pfeffer, 2007; Lin, Kraus, Wilkenfeld, & Barry, 2008). Generally speaking, we may expect that artificial agents may similarly be more successful when arguing with human users, if they can anticipate human reactions to various abstract argumentation frameworks. With that goal in mind, our results suggest that artificial agents may be better off avoiding discussion that may reveal a defeater, even if the agent has a counter-argument to that defeater; but should be ready to use floating reinstatement as well as simple reinstatement in order to neutralise a defeater raised by the human user. These kinds of heuristics can be incorporated into a decision-theoretic model of a persuasive agent that interacts with users using natural language (Reed, 1998; Grasso, Cawsey, & Jones, 2000). Such agents may also be complemented by domain-specific knowledge of effective argumentation strategies (e.g. in the domain of genetic counselling (Green, 2007) or healthy diet promotion (Mazzotta, Rosis, & Carofiglio, 2007)). Going beyond our specific results, by building up a corpus of argument structures and how they are evaluated, it may be possible to use machine learning techniques to build models that predict how people will react to novel argument structures.

Independently of our specific results, we hope to have convinced the reader that the wealth of scientific methodology from psychology can give a new perspective on the problems raised when formalising argumentation and developing argument evaluation semantics. We hope that our claims and findings can prompt researchers working on the computational modelling of argument to explore new avenues of investigation inspired by, and validated against, empirical evidence from psychology and cognitive science.

We also hope to have excited cognitive scientists working on human reasoning about the growing literature on formal models of argumentation. These models, and their associated normative properties, have great potential in complementing existing research on human reasoning, and providing conceptual means for dealing with highly complex inference structures.

References

Amgoud, L., Bonnefon, J. F., & Prade, H. (2005). An argumentation-based approach for multiple criteria decision. *Lecture Notes in Computer Science*, *3571*, 269–280.

Baroni, P., & Giacomin, M. (2007). On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, *171*(10–15), 675–700.

Baroni, P., Giacomin, M., & Guida, G. (2005). SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, *168*(1–2), 162–210.

Bench-Capon, T. J. M., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, *171*(10–15), 619–641.

Benferhat, S., Bonnefon, J.-F., & da Silva Neves, R. (2005). An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese*, *146*(1-2), 53-70.

Benthem, J. van. (2008). Logic and reasoning: do the facts matter? *Studia Logica*, *88*(1), 67-84.

Besnard, P., & Hunter, A. (2008). *Elements of argumentation*. Cambridge MA, USA: MIT Press.

Bonnefon, J. F. (2004). Reinstatement, floating conclusions, and the credulity of mental model reasoning. *Cognitive Science*, *28*, 621–631.

Bonnefon, J. F. (2009). A theory of utility conditionals: Paralogical reasoning from decision-theoretic leakage. *Psychological Review*, *116*, 888–907.

Bonnefon, J. F., Da Silva Neves, R. M., Dubois, D., & Prade, H. (2008). Predicting causality ascriptions from background knowledge: Model and experimental validation. *International Journal of Approximate Reasoning*, *48*, 752–765.

Bonnefon, J. F., Dubois, D., Fargier, H., & Leblois, S. (2008). Qualitative heuristics for balancing the pros and cons. *Theory and Decision*, *65*, 71–95.

Caminada, M., & Amgoud, L. (2007). On the evaluation of argumentation formalisms. *Artificial Intelligence*, *171*, 286-310.

Caminada, M. W. A. (2006a). On the issue of reinstatement in argumentation. In M. Fisher, W. van der Hoek, B. Konev, & A. Lisitsa (Eds.), *Logics in Artificial Intelligence, 10th European Conference, JELIA 2006, Liverpool, UK, September 13-15, 2006, Proceedings* (Vol. 4160, p. 111-123). Springer.

Caminada, M. W. A. (2006b). Semi-stable semantics. In P. Dunne & T. Bench-Capon (Eds.), *Proceedings of the 1st international conference on computational models of argument (comma)* (pp. 121–130). Amsterdam, Nethrelands: IOS Press.

Da Silva Neves, R. M., Bonnefon, J. F., & Raufaste, E. (2002). An empirical test for patterns of nonmonotonic inference. *Annals of Mathematics and Artificial Intelligence*, *34*, 107–130.

Dubois, D., Fargier, H., & Bonnefon, J. F. (2008). On the qualitative comparison of decisions having positive and negative features. *Journal of Artificial Intelligence Research*, *32*, 385–417.

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

Evans, J. S. B. T., & Over, D. E. (2004). *If.* Oxford: Oxford University Press.

Ford, M. (2004). System LS: A three tiered nonmonotonic reasoning system. *Computational Intelligence*, *20*, 89–108.

Ford, M., & Billington, D. (2000). Strategies in human nonmonotonic reasoning. *Computational Intelligence*, *16*, 446–468.

Gal, Y., & Pfeffer, A. (2007). Modeling reciprocity in human bilateral negotiation. In *National conference on artificial intelligence (aaai).* Vancouver, British Columbia.

Grasso, F., Cawsey, A., & Jones, R. (2000). Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, *53*(6), 1077–1115.

Green, N. (2007). A study of argumentation in a causal probabilistic humanistic domain: Genetic counseling. *International Journal of Intelligent Systems*, *22*, 71-93.

Horty, J. F. (2002). Skepticism and floating conclusions. *Artificial Intelligence*, *135*(1-2), 55-72.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646Ű678.

Lin, R., Kraus, S., Wilkenfeld, J., & Barry, J. (2008). Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence*, (accepted).

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

Pelletier, F. J., & Elio, R. (1997). What should default reasoning be, by default? *Computational Intelligence*, *13*(2), 165-187.

Pelletier, F. J., & Elio, R. (2005). The case for psychologism in default and inheritance reasoning. *Synthese*, *146*(1-2), 7-35.

Pfeifer, N., & Kleiter, G. D. (2005). Coherence and nonmonotonicity in human nonmonotonic reasoning. *Synthese*, *146*, 93–109.

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*, 206–217.

Pietarinen, A.-V. (2003). What do epistemic logic and cognitive science have to do with each other? *Cognitive Systems Research*, *4*(3), 169-190.

Politzer, G., & Bonnefon, J.-F. (2006). Two varieties of conditionals and two kinds of defeaters help reveal two fundamental types of reasoning. *Mind & Language*, *21*(4), 484–503.

Prakken, H. (2002). Intuitions and the modelling of defeasible reasoning: some case studies. In S. Benferhat & E. Giunchiglia (Eds.), *Proceedings of the 9th international workshop on nonmonotonic reasoning* (p. 91Ű102).

Rahwan, I., & McBurney, P. (2007). Guest editors' introduction: Argumentation technology. *IEEE Intelligent Systems*, *22*(6), 21–23.

Rahwan, I., & Simari, G. R. (Eds.). (2009). *Argumentation in artificial intelligence*. Springer.

Reed, C. (1998). *Generating arguments in natural language*. Unpublished doctoral dissertation, University College London.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*, 491–517.

Stenning, K., & Lambalgen, M. van. (2008). *Human reasoning and cognitive science*. Cambridge MA, USA: MIT Press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453-458.

van Eemeren, F. H., Grootendorst, R. F., & Henkemans, F. S. (Eds.). (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary applications*. Mahwah NJ, USA: Lawrence Erlbaum Associates.

Appendix A
Materials used in Study 1

Argument 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.

Argument Set 2

(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.

Argument 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.

Argument Set 4

(A) John has no way to know Leila's password. Therefore, Leila's emails 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.

Argument 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.

Argument Set 6

(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.

Appendix B
Materials used in Study 2

Argument Set 1

(A) Cody does not fly. Therefore, Cody is unable to escape by flying.

(B) Cody is a bird. Therefore, Cody flies.

(C) Cody is a rabbit. Therefore, Cody is not a bird.

(D) Cody is a cat. Therefore, Cody is not a bird.

Argument Set 2

(A) Smith does not follow American spelling. Therefore, Smith writes 'colour' instead of 'color'.

(B) Smith speaks American English. Therefore, Smith follows American spelling.

(C) Smith was born and brought up in England. Therefore, does not speak American English.

(D) Smith was born and brought up in Australia. Therefore, does not speak American English .

Argument Set 3

(A) The car did not slow down. Therefore, the car approached the signal at the same speed or higher.

(B) Louis applied the brake. Therefore, the car slowed down.

(C) Louis applied the accelerator instead. Therefore, Louis did not apply the brake.

(D) Louis applied the clutch instead. Therefore, Louis did not apply the brake.

Argument Set 4

(A) Stephen is not guilty. Therefore, Stephen is to be free from conviction.

(B) Stephen was seen at the crime scene at the time of the crime. Therefore, Stephen is guilty.

(C) Stephen was having dinner with his family at the time of crime. Therefore, Stephen was not seen at the crime scene at the time of the crime.

(D) Stephen was watching football with his friends in the stadium at the time of the crime. Therefore, Stephen was not seen at the crime at the time of the crime.