Utrecht University

Artificial Intelligence master

Empirical Evaluation of Simple Reinstatement in Formal Models of Argumentation

Elfia Bezou-Vrakatseli
6647782

Supervisors: Henry Prakken
Chris Janssen

January 22, 2021

Table of Contents

Abstract

This paper presents an empirical study of simple reinstatement, which purports to replicate existing findings on imperfect reinstatement and to investigate potential explanations, such as the disruption of the suspension of disbelief and a variation of it. The motivation behind it derives from the importance of replication studies and from the significance of the interpretation of this empirical finding for argumentation, as imperfect reinstatement is often used as supporting evidence in today's work on graded acceptability. To this end, an experiment was designed where people evaluated the conclusion of an argument. The ratings from receiving arguments sequentially were compared against ratings given under two alternative manners of argument presentation: receiving all arguments at once and receiving them sequentially after being introduced to the relevant theory. A significant effect of the latter was found on ratings, but not of the former. The results successfully replicated the existing findings but did not validate the two aforementioned explanations. In order to decide whether said findings can be interpreted as supporting evidence for graded semantics, alternative explanations need to be investigated first (such as directionality of attacks and order of arguments), along with whether these findings can be generalized to fit the wider spectrum of the theoretical features they seem to instantiate.

Introduction

The formal study of argumentation is a very fertile area in artificial intelligence; both humans and AI agents use argumentation for their internal reasoning and for their interaction with other agents. Formal models of Argumentation are making remarkable contributions to AI (for example, defining semantics of logic programs, implementing persuasive medical diagnostic systems, studying negotiation dialogs in multi-agent systems), and AI has helped Argumentation Theory grow—for instance, it has generated appropriate, formal tools for argument analysis, evaluation, and visualisation (Rahwan & Simari, 2009).

This project constitutes an empirical evaluation of formal models of argumentation, by comparing their outcomes with answers given by humans. The use of experiments as a tool for validating argumentation models can be highly useful because, despite the fact that they are meant to be normative, said models are also meant to formalise how humans actually reason; the closer argumentation models are to actual reasoning, the more easily humans (or AI agents) can follow their norms.

Argumentation consists of two major branches; abstract argumentation theory, introduced by Dung (1995), where "one models arguments by abstracting away from their internal structure to focus on the relations of conflict between them" (Cramer & Guillaume, 2018), and structured argumentation theory (Besnard et al., 2014), where "one additionally models the internal structure of arguments through a formal language in which arguments and counterarguments are constructed" (Cramer & Guillaume, 2018). In the present study we will see how the use of a structured account can enhance existing studies on argumentation, which use abstract accounts of argumentation, by giving helpful insights and providing alternative explanations.

In particular, this study purports to shed some new light on the findings of Rahwan et al. (2010) on simple reinstatement. Rahwan et al. (2010) found that humans' confidence in an argument drops once an attacker is introduced, even after defending/reinstating said argument (i.e., even after the introduction of a counter-attacker). These findings are in the middle of the current discourse of argumentation, regarding graded acceptability and its corresponding semantics. The present study collects behavioural data about the way humans reason and constitutes, firstly, a replication study of Rahwan et al. (2010) and, secondly, investigates various explanations of said findings, and whether they can be regarded as supporting evidence of graded acceptability.

The importance of replicating results of past studies has been increasingly rising, as the *replication crisis* keeps growing. The replication crisis is a big, currently ongoing, methodological crisis in many scientific fields and, particularly, in the social sciences. Studies are more and more often found to be impossible (or very difficult) to replicate or reproduce, a problem that was addressed by Kahneman (2012). Of course, as the reproducibility of experimental results is an essential part of the scientific method, such inability can have grave consequences for many fields of science. Since the findings of Rahwan et al. (2010) have become central in the current work on graded semantics, they carry great importance, thus it is vital to retest them.

Graded acceptability was introduced by Cayrol & Lagasquie-Schiex (2005), who, building on Dung's (1995) framework, proposed the notion of *graduality* in the process of argument selection. The first study on graded, or gradual, acceptability that directly referred to Rahwan et al. (2010) as supporting evidence is by Grossi & Modgil (2015). They also introduced a generalisation of Dung's notion of acceptability that takes into consideration the numbers of attacks and counterattacks on arguments. Their theory "can arbitrate between competing preferred extensions and captures a specific form of accrual in instantiated argumentation", building on the assumptions that, all else being equal, "having fewer attackers

is ‘better’ than having more” and “having more defenders is ‘better’ than having fewer” (*ibid*), which are consistent with Dung’s (1995) semantics. They also introduce a third assumption that between the two, having fewer attackers takes precedence over having more defenders.

Another relevant study on graded acceptability is by Amgoud & Ben-Naim (2013), who propose a semantics with two main features: “an attack weakens its target but does not kill it” and “the number of attackers has a great impact on the acceptability of an argument”. Their semantics is based on four graded considerations: weakening, counting, relativity, and graduality. In essence, Amgoud & Ben-Naim (2013; 2016) formalise through postulates semantics of graded acceptability, including the aforementioned principle that it is better having fewer attackers than more (Grossi & Modgil, 2015; Grossi & Modgil, 2019), as well as the assumption that Rahwan et al.’s (2010) findings seem to support that, when attacked, an argument is not entirely killed, even its attacker is unattacked.

As already mentioned, the study of Rahwan et al. (2010) has become central in the current argumentation discourse, since their findings are seen as supporting evidence of graded acceptability. Imperfect reinstatement is being interpreted as the result of people applying the normative principle “all else being equal, having fewer attackers is ‘better’ than having more”, a possible explanation that was not considered/discussed by Rahwan et al. (2010) in their paper and that is argued against by Prakken & de Winter (2018). Overall, such an interpretation requires overcoming three hurdles: (1) ruling out all alternative explanations, (2) confirming that these findings can generalize to fit the wider spectrum of the theoretical features they (are supposed to) instantiate, and (3) accepting the assumption that people reason rationally and that empirical findings have relevance to normative theories.

This study consists of an experiment that purports to (1) replicate the results of Rahwan et al. (2010) and (2) examine whether these experimental results can confirm the aforementioned principle, or whether the criticisms of Prakken & de Winter (2018) are justified. For the latter, the experiment examines the effect of the manner of argument presentation on people’s confidence in an argument’s conclusion, in order to test the validity of two potential explanations of imperfect reinstatement: (1) the disruption of the suspension of disbelief (i.e., the suggested explanation by Rahwan et al., 2010) and (2) its variation by Prakken & de Winter (2018). Moreover, two other possible explanations are discussed: (1) the directionality of attacks and (2) the order of the arguments. The former, and more important, explanation derives from the use of a structured account, illustrating the limitations of a purely abstract approach.

Preliminaries

This section provides a brief introduction to formal (abstract and structured) argumentation, introducing some of its key elements. In abstract argumentation, arguments are modelled by abstracting away from their internal structure, whereas in structured argumentation theory, the internal structure of arguments is also modelled through a formal language. The technical background of both argumentation theories can be found below, where some of their main definitions and patterns are presented.

Abstract Argumentation

Abstract argumentation focuses on conflict relations between arguments. An *abstract argumentation framework* (*AF*) is a pair $\langle A, attack \rangle$, where A is a set of arguments and $attack \subseteq A \times A$ is a (binary) defeat relation. A significant remark is that, unlike classical logic, an attack of an argument and its counterargument can be asymmetric (i.e., an argument A can attack an argument B without B attacking A). The theory about *AF*s was initiated by Dung (1995), and below its most important points are summarised.

Status assignments

- The status assignment (SA) of *AF* assigns to (zero or more) members of *A* either the status *In* or the status *Out* (but not both) such that:
    1. An argument is *In* iff all arguments that defeat it are *Out*.
    2. An argument is *Out* iff some argument that defeats it is *In*.
- An argument is *Undecided* if it is neither *In* nor *Out*: $Undecided = A / (In \cup Out)$.
- A status assignment is *stable* iff $Undecided = \varnothing$ (stable semantics).
- A status assignment is *preferred* iff *In* is $\subseteq$-maximal (preferred semantics).
- A status assignment is *grounded* iff *In* is $\subseteq$-minimal (grounded semantics).

Justification status of arguments

1. Grounded semantics:
    - *A* is justified iff *A* is *In* in the grounded SA.
    - *A* is overruled iff *A* is *Out* in the grounded SA.
    - *A* is defensible iff *A* is *Undecided* in the grounded SA.
2. Preferred/stable semantics:
    - *A* is justified iff *A* is *In* in all preferred SA.
    - *A* is overruled iff *A* is *Out* in all preferred SA.

- *A* is defensible iff *A* is *In* in some but not all preferred SA.

Extensions

The theory also identifies sets of arguments (i.e., *extensions*) which are internally coherent and defend themselves against attack.

- An argument a ∈ *A* is *defended* by a set S ⊆ *A* (or is acceptable with respect to it) if for all b ∈ *A*: if b attacks a, then some c ∈ S attacks b.
  - S *defends A* if all defeaters of *A* are defeated by S.
- A set S of arguments is *conflict-free* iff no argument in S defeats an argument in S.
- Relative to a given *AF*, E ⊆ *A* is *admissible* if E is conflict-free and defends all its members.
- E is a *preferred* extension if E is a ⊆-maximal admissible set.
- E is a *stable* extension if E is admissible and attacks all arguments outside it.
- E ⊆ *A* is the *grounded* extension if E is the least fix-point of operator F, where F(S) returns all arguments defended by S.
- E is a *complete extension* if E is admissible and *A* ∈ E iff *A* is defended by E; any preferred, stable or grounded extension is a complete extension.
- For T ∈ {complete, preferred, grounded, stable}, X is *sceptically* justified under the T semantics if X belongs to all T extensions.
- For T ∈ {complete, preferred, grounded, stable}, X is *credulously justified* or under the T semantics if X belongs to at least one T extension.

An argumentation framework can be represented as a directed graph, where nodes are complete arguments (i.e., a premise and a conclusion) and the directed arrows represent defeats amongst arguments. In the following graph, *simple reinstatement* is illustrated:

Figure 1: A graph of simple reinstatement
Argument A is defeated by argument B, which is defeated by argument C.
A and C are sceptically justified.

Example

Let us consider three arguments A, B, and C, such that B defeats A and C defeats B, as illustrated in Figure 1.

A: “The apple on the table is red, because I see it red.”

B: “Mary is standing closer to the apple and she says it’s green, so it’s green”

C: “Mary is colourblind, so the apple is red.”

Structured Argumentation

Structured argumentation defines the notion of an abstract argumentation system as a structure. Structured argumentation is characterised by the family of ASPIC-like frameworks, such as the ASPIC+ framework (Modgil & Prakken, 2012), assumption-based argumentation (ABA) (Dung, Kowalski & Toni, 2006), etc. Below some key definitions are presented.

Argumentation System

An argumentation system (*AS*) is a tuple $AS = (L, R, -, n)$ where:

- L is a logical language consisting of propositional or ground predicate-logic literals.
- $R = R_S \cup R_d$ is a set of strict (R_s) and defeasible (R_d) inference rules of the form $\varphi_1,\ldots, \varphi_n \rightarrow \varphi$ and $\varphi_1,\ldots, \varphi_n \Rightarrow \varphi$, respectively (where φ_i, φ are meta-variables ranging over well-formed formulas in L), such that $R_S \cap R_d = \emptyset$.
  - $\varphi_1,\ldots, \varphi_n$ are called the *antecedents* and ψ the *consequent* of the rule.
- $-$ is a function from L to 2^L, such that:
  - φ is a *contrary* of ψ if $\varphi \in \overline{\psi}$, $\psi \notin \overline{\varphi}$;
  - φ is a *contradictory* of ψ (denoted by ‘$\varphi = -\psi$’), if $\varphi \in \overline{\psi}$, $\psi \in \overline{\varphi}$.
- n is a partial function such that n: $R_d \rightarrow L$.

Knowledge Base

A knowledge base in an $AS = (L, R, n)$ is a set $K \subseteq L$ consisting of two disjoint subsets K_n (the axioms) and K_p (the ordinary premises).

Arguments

An argument A on the basis of a knowledge base K in an argumentation system AS is a structure obtainable by applying one or more of the following steps finitely many times:

1. φ if $\varphi \in K$ with: Prem(A) = {φ}; Conc(A) = φ; Sub(A) = {φ}; TopRule(A) = undefined.
2. $[A_1],\ldots,[A_n] \rightarrow \psi$[2] if $A_1,\ldots,A_n$ are arguments such that $\text{Conc}(A_1),\ldots,\text{Conc}(A_n) \rightarrow \psi \in R$ with:
   Prem(A) = $\text{Prem}(A_1) \cup \ldots \cup \text{Prem}(A_n)$, Conc($A$) = ψ, Sub(A) = $\text{Sub}(A_1) \cup \ldots \cup \text{Sub}(A_n) \cup \{A\}$, TopRule($A$) = $\text{Conc}(A_1),\ldots, \text{Conc}(A_n) \rightarrow \psi$.

3. $[A_1],\ldots,[A_n] \Rightarrow \psi$ if $A_1,\ldots,A_n$ are arguments such that $\text{Conc}(A_1),\ldots, \text{Conc}(A_n) \Rightarrow \psi \in R$ with: $\text{Prem}(A) = \text{Prem}(A_1) \cup \ldots \cup \text{Prem}(A_n)$, $\text{Conc}(A) = \psi$, $\text{Sub}(A) = \text{Sub}(A_1) \cup \ldots \cup \text{Sub}(A_n) \cup \{ A \}$, $\text{TopRule}(A) = \text{Conc}(A_1),\ldots, \text{Conc}(A_n) \Rightarrow \psi$.

Example

Let us consider a knowledge base in an argumentation system with L consisting of n_1, n_2, p_1, p_2, p_3, p_4, s, t, u, v and their negations, where p_4 and s are negating each other (with $K_n = \{n_1, n_2\}$ and $K_p = \{p_1, p_2, p_3, p_4\}$), with the defeasible rules $R_d = \{d_1, d_2, d_3, d_4\}$, where

n_1 = "A pandemic is extreme circumstances."

n_2 = "The right to privacy is part of the Universal Declaration of Human Rights of UN."

p_1 = "Lockdown requires surveillance."

p_2 = "In extreme circumstances rights can be restricted."

p_3 = "Surveillance constitutes violation of the right to privacy."

p_4 = "The human rights established by UDHR should not be violated."

s = "During a pandemic, certain rights can be restricted."

t = "The right to privacy should not be violated."

u = "Surveillance is bad."

v = "We should not enforce lockdown."

d_1: n_1 , p_2 => s

d_2: n_2 , p_4 => t

d_3: p_3, t => u

d_4: p_1, u => v

We also define the n function $n(d_i) = d_i$, which formally assigns well-formed formulas d_i (informally referred to as inference rules) to said rules. For instance, n (n_1 , p_2 => s) = d_1.

The deriving arguments are:

A_1: p_1

A_2: p_3

A_3: n_2

A_4: p_4

A_5: A_3, A_4 => t

A_6: A_2, A_5 => u

A_7: A_1, A_6 => v

B_1: n_1

B_2 : p_2

B_3: B_1, B_2 => s

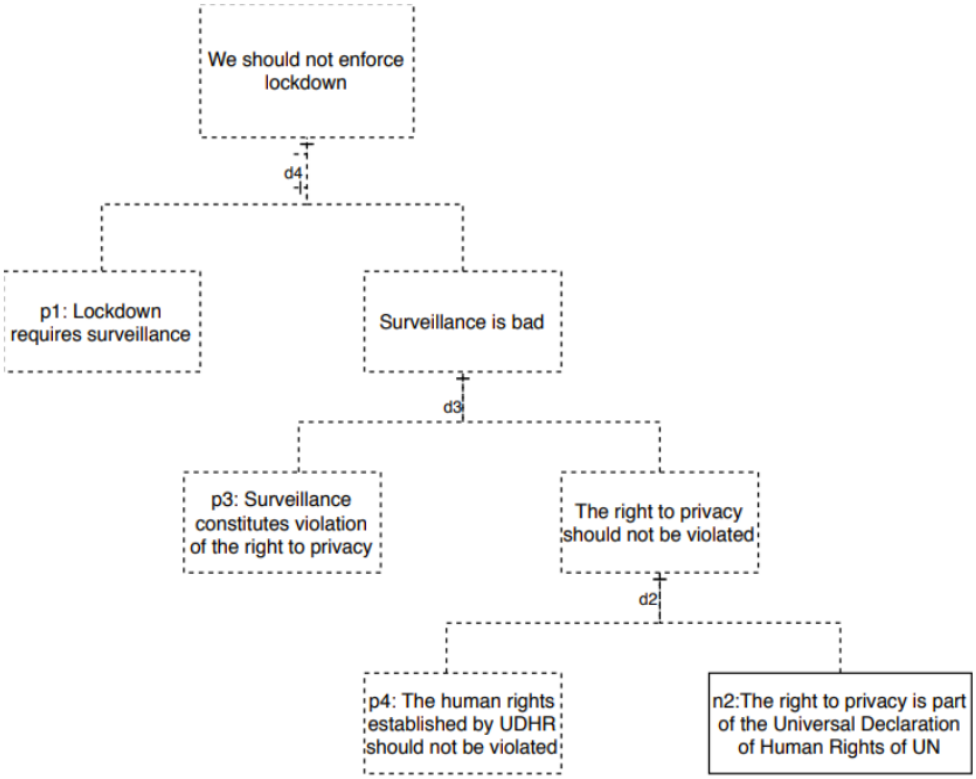

Figure 2: An argument for v

An argument for *v* is displayed in Figure 2, with Prem(A_7) = {p_1, p_3, p_4, n_2}, Conc(A_7) = *v*, DefRules(A_7) = {d_2, d_3, d_4}, TopRule(A_7) = d_4, Sub(A_7) = {A_1, A_2, A_3, A_4, A_5, A_6, A_7}. Another argument for *s* is depicted in Figure 3, with Prem(B_3) = {n_1, p_2}, Conc(B_3) = *s*, DefRules(B_3) = {d_1}, TopRule(B_3) = d_1 , Sub(B_3) = {B_1, B_2, B_3}.

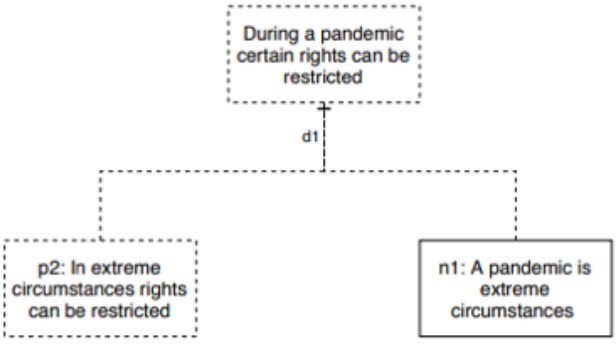

Figure 3: An argument for s

Attack

An argument A *attacks* an argument B iff A *undercuts* or *rebuts* or *undermines* B, where:

- A *undercuts* B (on B′) iff Conc(A) = −n(r) and B′ ∈ Sub(B) such that B′'s top rule *r* is defeasible (A attacks an inference of B).
- A *rebuts* B (on B′) iff Conc(A) = −φ for some B′ ∈ Sub(B) of the form $B''_1,\ldots,B''_n \Rightarrow \varphi$ (A attacks the conclusion of B).
- A *undermines* B (on *φ*) iff Conc(A) = −φ for some φ ∈ Prem(B) ∩ K_p (A attacks a premise of B).

An argument can be indirectly attacked when one of its proper sub-arguments is directly attacked.

Example

The attack relation indicates which arguments are in conflict. In our running example, B_3 and A_4 are in conflict, directly attacking each other; B_3 attacks directly A_4 and indirectly A_5, A_6, A_7. In return, A_4 attacks B_3 directly (see Figure 4). Moreover, since p_4 is an ordinary premise of A_7, B_3 undermines A_7 on A_4. Lastly, although no undercutting takes place in this example, both arguments could be undercut, since all the inference rules they contain are defeasible.

Directionality

When argument A attacks argument B and argument B attacks argument A, we have a symmetric attack (bidirectional conflict), like in the case of B_3 and A_4. In structured argumentation theory, as in abstract argumentation, an attack between an argument and its counterargument can also be non-symmetric (a unidirectional attack); one argument can attack another, without the latter attacking the former (for example, when argument A undercuts argument B). Lastly, the ASPIC+ framework allows to specify a preference ordering between the defeasible premises and rules, which gives rise to a preference order between arguments.

An attack in ASPIC+ from A to B can be unidirectional either if A undercuts B or as the result of a preference ordering (for instance, some subargument of A is strictly preferred to some subargument of B). However, in ABA, an attack from A to B is actually based on an attack of the conclusion of A to an assumption of B, where assumptions roughly correspond to premises in ASPIC+. Thus, attacks in ABA correspond roughly to the undermings in ASPIC+, which, unlike in ASPIC+, can be unidirectional even without orderings.

Structured Argumentation Frameworks

Let *AT* be an argumentation theory, i.e. a pair (*AS*, *K*). A structured argumentation framework (*SAF*) defined by *AT*, is a triple $\langle A, C, \preceq \rangle$, where *A* is the set of all arguments on the basis of *K* in *AS*, $\preceq$ is an ordering on *A*, and (X, Y) $\in$ *C* iff *X* attacks *Y*.

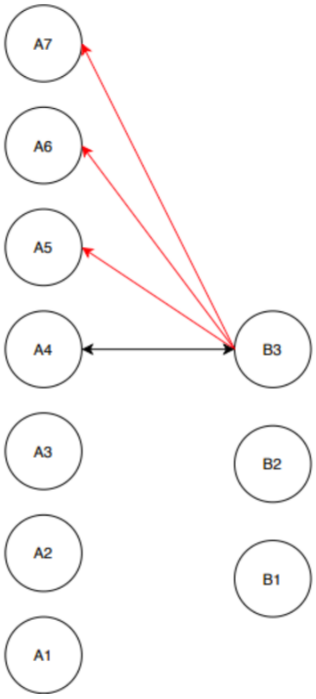

Figure 4: Attacks
Direct attacks are depicted in black, while indirect are in red.
A_4 and B_3 are in a bidirectional conflict (symmetric attack).

Defeat

A defeats B iff either A undercuts B; A rebuts or undermines B on B' and A$\not\prec$B'. Undermings and rebuttals are successful only if the attacked argument is not preferred over its attacker.

Example

To figure out the successful attack, or defeat, one needs to compare the arguments that are in a direct rebutting or undermining relation. In our running example there are no orderings, thus B_3 and A_4 both defeat each other.

Argumentation frameworks

Abstract argumentation frameworks are generated from *SAF*s by letting the attacks from an *AF* be the defeats from a *SAF*. An abstract argumentation framework (*AF*) corresponding to a *SAF* = $\langle A, C, \preceq \rangle$ (where *C* is ASPIC+'s attack relation) is a pair $(A, attack)$ such that *attack* is the defeat relation on *A* determined by SAF.

The need of empirical evaluation

As already mentioned, this study consists of an empirical evaluation of cases of simple reinstatement. Thus, it is important at this point to highlight the significance of an empirical evaluation for the quality and adequacy of formal models. Since formal models are usually meant to be normative (instead of descriptive), the following question is engendered: why are we in need of their empirical evaluation?

Let us begin by the distinction between normative and descriptive theories. A normative theory consists in a "systematic set of claims about how argumentation *should* be conducted" (Jackson, 1988) and, thus, when a formal argumentation model is (meant to be) normative, it constitutes an idealised picture of argumentation. On the other hand, a descriptive theory consists in "a systematic set of claims about argumentation practice" (*ibid*), and the goal of a descriptive model is to picture argumentation as it happens. The two differ greatly, thus, since the first represents how people *ought* to be arguing, whereas the second pictures how human argumentation actually *is*.

Thus, why do we need to know how people actually argue? One could argue that epistemic rules are supposed to be formulated independently of psychology since they are normative. However, the answer to the aforementioned question can be found in the very purpose of such theories; why one needs an argumentation theory in the first place. A central challenge for an argumentation theory is to give an account for argumentation occurring under non-ideal circumstances, by non-ideal participants (van Eemeren et al., 2002). Amgoud & Ben-Naim (2016) define argumentation as a *social* activity purporting to "increase (or decrease) the acceptability of a given standpoint for an audience by putting forward arguments". Thus, overall, are formal models useful, no matter how far from the manner humans actually reason?

Perhaps epistemic rules are normative, but correct epistemic norms must also be "reason-guiding" (Pollock, 1986), meaning that they are supposed to guide humans in their reasoning and belief-formation process. Formal models can provide us with a normative standard against which practice can be compared (van Eemeren et al., 2002); the more 'everyday argumentative discourse' and formal models of argumentation remain in close proximity, the more they can help each other. Even if the goal is making 'real' argumentation look as much as possible like 'ideal', the latter should be relevant to the former, since, otherwise, such an attempt is futile. Therefore, although normative, argumentation models are (and should remain) close to how humans actually reason, so that humans can more easily follow the norms of said models.

Van Benthem (2008) advocates the benefits of enriching theories in ways that they approach facts, listing various examples where logic has been able to absorb insights of ways people reason into making richer systems. Coining the term "New Psychologism", he describes how evidence obtained by systematic experimental design can enhance formal theories. According to him, since actual behaviour is influenced by abstract theory "through the design of new intelligent practices allowing for 'successful insertions' of behaviour into human lives" (van Benthem, 2008), the usual normative/descriptive distinction does not suffice.

The integration of human and machine reasoning is necessary for practical applications to usefully deploy logic-based reasoning techniques. This is also the reason that Dung's theory is as successful; it is founded on both computational and human argumentation (Modgil et al., 2013). In fact, computational models of argumentation can provide support for complex decision-making activities exactly because their semantics is in close accordance with human intuition (Cerutti et al., 2014). This applicability of argumentation theories, which should serve as reasoning guides, is also highlighted by Prakken (2020), where the benefits of an empirical (or "quasi-empirical") validation are discussed. Such validations bear the ability of developing, or enriching, normative theories in ways that are applicable by humans. Even if normatively adequate, a theory very foreign to people will end up not being applied, which means it will not be able to serve as useful guidance for human reasoning.

Although Prakken (2020) advocates that to fully validate abstract accounts of argumentation we need "instantiation with more concrete theories of reasoning and argumentation that have been themselves sufficiently validated" (i.e., theory-based validation), he suggests that empirical validation can be used at all levels of abstraction as a *partial* validation method. What is more, the fact that theories (might) deviate from the way humans argue in reality should not discourage experimental research or empirical validation. The goal is to state theories "in terms that are natural to people" (*ibid*); this way we increase the chances that they are applied by humans and, thus, theories can offer guidance for human reasoning.

A pragmatic approach to evaluating argumentation models, as described by Veltman (1987), advocates that a theory's ratification is about clarifying intuitions, rather than predicting them. This approach does not assume a pre-existing logic in argumentation; between two competing theories, "the better theory is the one so well motivated that people are prepared to allow it to guide their judgments whenever their intuitions leave them groping, and even to correct their judgments whenever their intuitions do not—not yet!—match the theory's predictions" (Veltman, 1987). A pragmatist views language as an instrument of communication and their main concern is ensuring that following their advice will make said instrument more useful. One should perceive, and make use of, the findings of the experimental procedure in

the appropriate manner; if a participant's judgement disagrees with the expected result of the theory, one should take into consideration the factors that led to it.

Let us now examine ways in which a descriptive study of argumentative practises can contribute to a normative theory. Firstly, an empirical evaluation of existing formal models can help us decide among them when they disagree. Moreover, even if they agree, they can offer insights to them, by highlighting matters with important theoretical implications, such as the social context of argumentation, the existence of implicit normative models, the justification of alternative models, and ways that normative models contribute to practice (Jackson, 1988). Besides, one must keep in mind that human reasoning is also governed by internalized rules (Pollock, 1986). People know how to reason (or have certain, personal ways to reason) and any descriptive attempt constitutes a "reconstruction of people's *own* normative ideas" (Jackson, 1988), i.e., ideas about whether to defend a statement and how to argue.

According to methodological descriptivism, "a theory can and should be tested by comparing what it has to say about the validity of the arguments it covers with the intuitive judgments of those who use the language concerned" (Veltman, 1987). A similar approach was followed by Rahwan et al. (2010) for the evaluation of formal argumentation models (which is more elaborately discussed in the following section). They proposed a *descriptive-experimental* approach, since they found limiting to only investigate through a normative perspective of argumentation. Rahwan et al. (2010) speak of an "interplay between logic and cognition", arguing that empirical findings from cognitive science have ameliorated epistemic logic and that computational argumentation can benefit from similar studies. Polberg & Hunter (2018), who used experiments with humans to evaluate certain frameworks, put it like this: "[w]ithout empirical evidence, we can accidentally increase the gap between applying argumentation and successfully applying argumentation in real life situations".

As argumentation theory is a quite pregnant area in AI, knowing how humans argue can, for instance, contribute to the design of software agents that argue with humans, or provide reliable support to human evaluation of arguments (Rahwan et al., 2010). Rahwan et al. (2010) postulate that being able to anticipate human reactions, even in abstract argumentation frameworks, enhances the agents' arguing abilities. They list various studies where artificial agents have achieved better negotiation results with human users when following rational strategies derived from human behavioural data, instead of normative equilibrium strategies. Their descriptive-experimental approach is followed in this study as well; an approach that consists of methods of experimental psychology for studying argument-based reasoning by comparing the experimental results with the predictions that formal models of argumentation theories make.

Related Work

This section consists of previous studies which have conducted experiments about the way humans reason in order to evaluate argumentation models, or enrich argumentation theories.

Rahwan et al. (2010)

The study of Rahwan et al. is the first study that, "empirically investigated the cognitive plausibility of the formalisms of argumentation theory" (Cramer & Guillaume, 2018). It is still a central element of argumentation discourse and it constitutes the main motivation of this study.

Rahwan et al. (2010) focused on reinstatement (simple and floating). In some cases (such as simple reinstatement), preferred and grounded semantics are in agreement, but in other (such as floating reinstatement), the two semantics have different takes for argument acceptability. In the latter case, where there is a conflict between predictions, their study proposes the use of a *descriptive-experimental* approach, which consists of methods of experimental psychology, where human answers are compared to theoretical predictions. The descriptive-experimental approach is proposed because the authors believe it can provide useful insights to inform current (and future) semantics and, also, contribute to the design of software agents that argue with humans or assist humans with argument evaluation.

The study consisted of two separate studies; study1 investigated simple reinstatement and study2 floating reinstatement. I will elaborate on the former, as it is the object of this project. The first study followed a three-level (Base, Defeated, Reinstated), six-measure, within-subjects design. It consisted of six argumentative sets (see Appendix B1, sets 1-6), of three statements each; for simple reinstatement, each set consisted of statement A, statement B attacking A, and statement C attacking B (the type of attack corresponding to undermining in ASPIC+). The dialogue was divided into three stages, where first A, then A and B, and lastly all A, B and C were presented. The participants evaluated the Base (argument A), Defeated (arguments A and B), and Reinstated (arguments A, B, and C) stage sequentially, in this order.

The sets were short, simple, and, because of the neutral subject matter of the sentences used, not emotionally engaging to the participants, or likely to be affected by subjective views of the participants. Confidence in the conclusion was used as the dependent variable as, for each problem, the participants had to evaluate the conclusion of statement A by choosing a value from a 7–point scale ranging from *Certainly false* to *Certainly true*.

The results revealed that the confidence in the conclusion of argument A decreases once A is defeated and increases when A is defended (i.e., when the defeater is itself attacked by a

reinstating argument), though it stays significantly lower than the initial state (prior to the defeat). The researchers concluded from this that the participants reasoned in a way that "reflected the formal notions of defeat and reinstatement" and in agreement with grounded and preferred semantics. However, neither semantics predicts that the reinstated argument does not fully recover from a defeat.

One possible explanation the authors give is in terms of *suspension of disbelief*, which is involved in reasoning experiments that use natural language materials. They claim that participants are capable of thinking (and might, in fact, think of) different kinds of objections to the presented arguments, but they suspend these objections, this disbelief in the argument, for the sake of the experiment. However, when one objection is presented by the experimenter, said suspension is disrupted. Thus, the aforementioned, private objections, which are not explicitly ruled out in the problem, end up influencing the way participants reason. Adding to this, Prakken & de Winters (2018) have suggested a variation of this explanation, advocating that, after being introduced to an attacker, a subject's degree of belief in other possible attackers increases as well because the very introduction of an attacker leads them to consider other possible objections.

Cerutti et al. (2014)

Cerutti et al. investigated the evaluation of their participants in various forms of reinstatement by making use of a logic-programming-based approach to structured argumentation proposed by Prakken & Sartor (1997). The study consisted of two texts (a base one and an extended one) on four topics (weather forecast, political debate, used car sale, and romantic relationship), on which the participants were provided with a questionnaire to assess the justification status of natural language statements. In the base case, two conflicting positions (A and B) were presented and, later, a third 'preference' statement, in favour of the second argument, was presented, determining a successful defeat of the first argument. In the extended case, a fourth statement was added, one that was reinstating the first argument (by either undercutting the preference argument, rebutting the second argument, or rebutting the preference argument).

Each participant was given one of the eight scenarios and was asked to decide between "position of person A is correct" (PA), "position of person B is correct" (PB), or "I cannot determine whether either of positions is correct" (PU). Moreover, the authors investigated the support for the preference statement. To this end, participants were asked to rate a number of statements in terms of relevance (for the conclusion) and agreement.

The results showed that the majority of participants in the base case scenarios agree with the second position (PB), whereas in the extended case they could not decide (PU), as

predicted. This suggests a correspondence between the formal theory introduced and its representation in natural language. Moreover, the authors concluded that people evaluate preference relations depending on the domain, possibly because participants use the preference to support conclusions and the acceptance of a preference is connected to the domain. In addition to this, personal knowledge of the participants was found to have influenced the results, as can be seen in the weather forecast scenario, where many participants explained their decision with the statement "All weather forecasts are notoriously inaccurate".

A significant limitation of this study is that (in the extended case) a different type of attack was used in each domain for the reinstatement of PA. Thus, in order to confirm whether the domain has a significant effect in the extended case as well, as found by the results of this study, the two factors (i.e., type of attack, domain) have to be investigated independently. Lastly, as each participant received only one text to evaluate (i.e., the base and extended scenarios were judged by different people), it has to be investigated whether similar results are achieved when the same person is shown both the former scenarios and the latter.

Cramer & Guillaume (2018)

Cramer and Guillaume tested whether the assumptions about the directionality of attacks between natural language arguments correspond to the manner humans actually evaluate said arguments. They pointed out that Rahwan et al. (2010) had not checked whether the way their participants perceived the directionality of the arguments' attacks was the same with the directionality they intended to present. In particular, Rahwan et al. (2010) assumed an asymmetric attack between their arguments, despite using underminings (which in ASPIC+ give rise to bidirectional attacks). Thus, Cramer & Guillaume (2018) investigated the way humans evaluate arguments and the way they perceive attacks' directionality.

The study consisted in two empirical cognitive studies; one with a group of naive adults and one with experts in formal argumentation theory. It addressed the questions "Do humans systematically interpret certain kinds of conflict in a non-symmetric, directed way?", and, if yes, "Is there any correspondence between the notion of directionality that humans employ when evaluating arguments and the criteria according to which structured argumentation frameworks like ASPIC+ and ABA determine the directionality of attacks?". In particular, they investigated whether humans interpret underminings as unidirectional (as also suggested by ABA), or whether humans' evaluation of directionality coincides with ASPIC+.

The authors used the four arguments of Rahwan et al.'s (2010) of the floating reinstatement case, thirty-six sets with the same structure (i.e., corresponding to the same abstract argumentation framework), and four sets of different structure and number of

arguments. They varied the nature of the attack type and the contexts of the arguments (for example, pet caring, scientific publications, etc.). In the naive adults' group, participants were shown two arguments at a time and were asked to decide on their acceptability status (whether they accept them, reject them, or consider them undecided) after being instructed to consider (both) non-conflicting arguments accepted, arguments in a symmetric conflict undecided, and in an asymmetric attack of A to B to reject the latter and accept the former. The experts' group received the whole argument sets and participants were asked to indicate all attack relations.

The results showed that some conflicts between arguments are systematically interpreted by humans as unidirectional attacks. Overall, the majority judgment coincided with the ASPIC+-based predictions. Specifically, in the case of 'simple undermining'—the one used in Rahwan et al. (2010)—the majority judged the attack as bidirectional (as suggested by ASPIC+), and not unidirectional (as suggested by ABA), rendering Rahwan et al.'s (2010) interpretation of their findings "problematic". However, a considerable minority of experts marked this attack as unidirectional, possibly suggesting that ABA framework has some reflection in the way (certain) experts judge simple undermining. Lastly, humans agreed with the predictions motivated by ASPIC+ to a varying degree, depending on the attack type, which could serve as an indication that, "the distinctions made by ASPIC+ are not fine-grained enough".

Cramer & Guillaume (2018, COMMA)

This empirical, cognitive study purported to compare different argumentation semantics to see which one best predicts human evaluation of arguments and whether the thematic context of the arguments plays a role. Two important novel features of this study are, firstly, that, in order not to take for granted any assumptions about the directionality of attacks between natural language arguments, two pilot studies were performed to test the perceived directionality of said attacks. Secondly, the study involved a group deliberation phase, in order to increase participants' performance in logical reasoning tasks.

Each participant was presented with a set of three to five natural language arguments, corresponding to the simple reinstatement, floating reinstatement or 3-cycle reinstatement argumentation framework, on three different thematic contexts (i.e., news reports, scientific publications, and precision of a calculation tool). The participants were asked to, firstly, draw the attack relation between the given arguments and, secondly, to judge the acceptability of each argument in light of the information provided in all arguments.

The results were in favour of the cognitive plausibility of Dung-style abstract argumentation theory, with CF2 and preferred semantics being better predictors for human argument acceptance than grounded. Moreover, in the cases in which all the standard argumentation semantics agree, humans evaluated the acceptability of arguments as predicted. Of course, there needs to be an investigation of whether these results are applicable to more generalized cases. Overall, their findings on simple reinstatement and floating reinstatement are in line with the findings of Rahwan et al. (2010).

Lastly, their findings also suggest an influence of participant's world knowledge on their decisions: in the case of an argument claiming that Donald Trump shot a lion, the predicted answer (i.e., 'accepted') was equally chosen as 'rejected'. World knowledge might have interfered either in the form of judging the *content* of the claim as highly implausible (thus favouring its rejection), or judging the *medium* providing the information (i.e., news reports were perceived as less reliable compared to scientific publications or calculation tools).

Cramer & Guillaume (2019)

This study of Cramer and Guillaume also consisted in an empirical, cognitive study that investigated human judgement on argument acceptability. It made use of complex argumentation frameworks since previous studies, such as Rahwan et al. (2010) and Cramer & Guillaume (2018), had been limited to small sets of simple argumentation frameworks (and thus some semantics could not be meaningfully distinguished). Cramer & Guillaume worked on improving Rahwan et al.'s (2010) methodology and applied this improved version on three different argumentation frameworks (adding also the 3-cycle reinstatement framework with five arguments). Also, they presented purely fictional scenarios (instead of referring to real-world entities and actions) to reduce the interference of participants' world knowledge, as they wanted to examine argument evaluation based on *attack relation*, rather than argument content.

As in many cases semantics made the same prediction, they focused on the instances where the two semantics under comparison differed. According to the results, grounded and CF2 semantics were found to be systematically better than the other semantics, in contrast with the previous study that preferred and CF2 were found to be the best predictors. The authors suggested that this difference is a result of the complexity of the argumentation frameworks used, which were, on top, represented with fictional scenarios. This indicates that the cognitive difficulty of the strategy is a factor in deciding which semantics best predicts participants' strategy (for instance, a cognitively more challenging task might lead to participants choosing a simplifying strategy, i.e., choosing 'undecided', more easily). Overall, results showed that for some participants (i.e., the ones choosing a cognitively simpler strategy) grounded

semantics is a better predictor, while for the other part of participants (that chose a more demanding strategy), their strategy is better predicted by CF2 semantics.

Moreover, the results suggested that the development of a novel semantics could be beneficial; one that behaves similarly to CF2 on most argumentation frameworks, but which "treats even cycles of length six or more in the way they are treated by preferred semantics". Lastly, one significant limitation highlighted by the authors is that semantics comparison only happens on the grounds of a single-outcome justification status, ignoring some of the information present in the full set of extensions provided by each semantics.

Polberg & Hunter (2018)

Polberg and Hunter empirically evaluated probabilistic argumentation frameworks (*PAFs*) and bipolar argumentation frameworks (*BAFs*). In Polberg & Hunter's (2018) experiment, participants were asked to judge dialogues in terms of agreement and structure. The authors claim that handling the uncertainty concerning both the participants' opinions about arguments and the structure of the graph is of critical importance, and that a Dung framework—equipped only with the standard semantics—is insufficient. They claim that gradual notions of acceptability are required along with a structure beyond attack, as *defence* (a type of a positive indirect relation between arguments, derived by direct attacks) does not account for all the possible forms of support between arguments. The authors clarified that their experiments are explorative, purporting to motivate future studies.

Their analysis consists in a comparison of their findings with constellation and epistemic approaches to probabilistic and bipolar argumentation. A probabilistic approach offers a more accurate user modelling, while an epistemic approach allows the representation of how much an argument is believed or disbelieved by a given person (i.e., more than three agreement states exist). The constellation approach allows the representation of the views of different people concerning the structure of the corresponding framework. Bipolar argumentation allows the expression of positive and negative relations between arguments.

Moreover, Polberg & Hunter (2018) empirically investigated prevailing assumptions in dialogical argumentation: that the involved parties correctly identify the intended statements posited by each other, realise all of the associated relations, conform to the three acceptability states, adjust their views in light of new, and correct information. The authors interpreted their results as supportive of the need of bipolar approaches and, also, as indications that people use their own personal knowledge to make judgments (without necessarily disclosing it), and that the introduction of new, correct information does not guarantee a change in one's beliefs.

Study: Simple Reinstatement

In this study the effect of the manner of argument presentation is examined. As mentioned in the previous section, Rahwan et al. (2010) found that, in simple reinstatement, belief in the conclusion of argument A decreases once A is defeated and increases when it is defended, but without re-reaching its initial level. I purport to investigate whether reinstatement is imperfect (i.e., whether the ratings in the reinstated stage are equally low) if the examples are presented to the subjects in different ways, as suggested in Prakken & de Winter (2018).

As mentioned in the Introduction, the current argumentation discourse revolves around the assertion that Rahwan et al.'s (2010) findings support certain features of several models of gradual argument acceptability. For instance, Grossi & Modgil (2015; 2019) have directly referred to said findings of imperfect reinstatement as supportive of their principle that "all else being equal having fewer attackers is 'better' than having more" and, consequently, as supportive of their graded semantics, which assigns higher status to unattacked arguments than to defended ones. Prakken & de Winter (2018) argued against such conclusions, advocating that there might be alternative reasons for these findings of unsuccessful reinstatement recovery and this study purports to investigate their criticisms.

As previously discussed, Rahwan et al. (2010) have suggested an explanation of their findings in terms of suspension of disbelief (see Rahwan et al., 2010). They advocated that the introduction of an attacker causes participants to stop suppressing their own, pre-existing objections, which were originally suppressed for the sake of the experiment (and which are not explicitly ruled out by the provided arguments). Following this suggestion, Prakken & de Winter (2018) have suggested a variation; the introduction of an attacker leads participants to consider alternative objections, increasing their degree of belief in other possible attackers, which are not explicitly ruled out in the case (i.e., the arguments) presented to them.

Both Rahwan et al.'s and Prakken & de Winter's suggestions are plausible, as, overall, the subject's prior/own beliefs are hard to eliminate completely, and it is almost impossible to ensure that someone is basing their confidence rating exclusively on the information provided by the given sentence, as suggested by various studies (for instance, Cerutti et al., 2014; Cramer & Guillaume, 2018; Evans & Over, 2004; Prakken & de Winter, 2018; Polberg & Hunter, 2018; Rahwan et al., 2010). Thus, if the arguments are presented to the subjects in different ways, different results may be derived.

To investigate this, three ways of argument presentation were examined in cases of simple reinstatement; the one used in Rahwan et al. (2010) and two variations. This study offers an experimental comparison between simple reinstatement cases where (1) arguments are

presented one-by-one (as in Rahwan et al., 2010); (2) all arguments are presented at once; (3) first all possible scenarios are presented—i.e., generalized forms of the arguments the participants encounter during evaluation—and then the arguments are presented one-by-one. Four hypotheses are tested.

1. When arguments are presented sequentially, like in Rahwan et al. (2010), participants' ratings for the conclusion of argument A in the reinstated stage are lower than in the base stage.

2. When all arguments are presented at once, 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).

3. When participants are first presented with all possible scenarios—i.e., when they are presented with generalized forms of the arguments they will encounter during evaluation, before evaluating them—and are then asked to evaluate the arguments:

    3.1. 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 case (where they have not seen all the possible scenarios beforehand).

    3.2. 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 case.

The first hypothesis merely predicts a successful replication of Rahwan et al.'s (2010) results. 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 changes the subject's belief in the initially presented argument even after it has been reinstated, it is possible that it is this very gradual process of presentation that influences the subject's degree of belief. To quote Rahwan et al. (2010), "[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, it is possible that if we eliminate the gradual factor of presentation, the suspension of disbelief will hold, as we can assume that said disruption results from the very gradual introduction of 'new evidence'.

Possibly, when an attacker is introduced after one has placed their confidence in an argument (or in the experimenter's argument, or in the experimenter), 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) in 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.

Extending on this thinking, and regarding my third and fourth hypothesis, we ought to consider another possible explanation and, thus, another manner of 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. The third and fourth hypothesis are taken from Prakken & de Winter (2018) 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 (and, thus, its conclusion untrue), 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 my first hypothesis constitutes a successful replication of Rahwan et al.'s (2010) findings, while a confirmation of the other three hypotheses underlines the importance of the way in which subjects are presented with arguments, proving it affects participants' confidence. Such confirmations would support the criticisms of Prakken & de Winter (2018), as, then, said findings could be interpreted as a result of the two aforementioned suggested explanations, and not as supportive evidence of graded argument acceptability.

Methods

The study followed a between-subjects design with three levels (receiving arguments sequentially, all at once, or being introduced with generalisations of all three arguments first and then receiving them sequentially). One third of the participants (i.e., the first group) received arguments in a sequential fashion of three stages: first argument A (base stage), then A and B (defeated stage), and lastly all A, B, and C (reinstated stage). The second third (i.e., the second group) received a text containing all three arguments. The third group received a text including generalisations of all three arguments first and then received the corresponding arguments in a sequential fashion, as did the participants of the first group. Since the first goal of my study is to replicate the results of Rahwan et al. (2010), an additional comparison between the ratings of the first group's participants was conducted (within-subjects). Overall, Rahwan et al.'s (2010) method was followed as closely as possible in terms of materials, procedure, and measurement.

Participants

The experimental sample consisted of 390 participants (130 for each group) within the age range of 18-65 years old. The average age of the first group was 30 years old, of the second group 33, and of the third group 28. Participation in the present study was voluntary, while the participants were recruited through personal contact. No strict inclusion criteria were defined, apart from participants being fluent in English, which was tested through a questionnaire in the beginning of the experiment (see Appendix A).

Design & Materials

Participants were asked to evaluate the acceptability status of natural language arguments by means of an online survey/questionnaire, which ran on Qualtrics (https://www.qualtrics.com/). The experimental sample was divided into three separate groups and, thus, three different surveys were created (one for each group), each corresponding to a different manner of argument presentation (which are described in detail in the following section). Participants were presented with exclusively one case/survey (between-subjects design).

The study consisted of a total of eight sets of arguments, consisting of three arguments each (argument A, argument B attacking A, and argument C attacking B, as depicted in Figure 1), with the attack relation between them corresponding to undermining in ASPIC+ (or 'simple undermining', as referred to by Cramer & Guillaume, 2018). The argument sets were adapted from Rahwan et al. (2010); six of the eight sets are the ones used in their experiment for simple

reinstatement and the other two sets were created by me, in the same structural line (i.e., with the same attack type and relations). No orderings were specified. The content of the argument sets can be found in Appendix B1. The complimentary text used for the third group can be found in Appendix B2 and it consists of the generalized forms of the aforementioned arguments (note that the structure of the generalized sentences is not identical to the original arguments, in order for the task not to resemble a memory task).

Each participant was presented with a questionnaire whose first pages consisted of personal questions and questions regarding their language capacity (see Appendix A). The next pages consisted of a brief example of an argument and the instructions about the procedure, where it was explicitly emphasized that participants had to stick solely to the provided information for evaluating the conclusion of argument A. The last part of the questionnaire consisted of the argument sets to be evaluated. 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 Figures 5a, 5b, 5c, 6, the parts of the survey for each stage of the first argument set are illustrated, as an example. Each stage was presented on a different page and no 'go back' option was available.

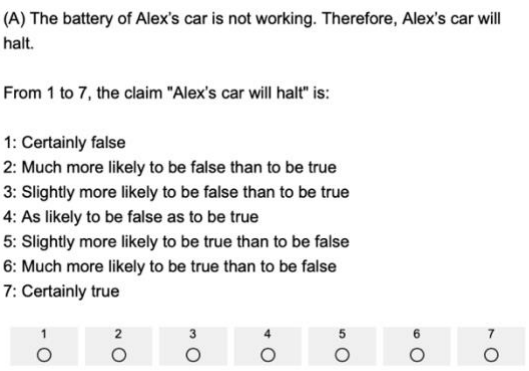

Figure 5a: Base stage of set 1

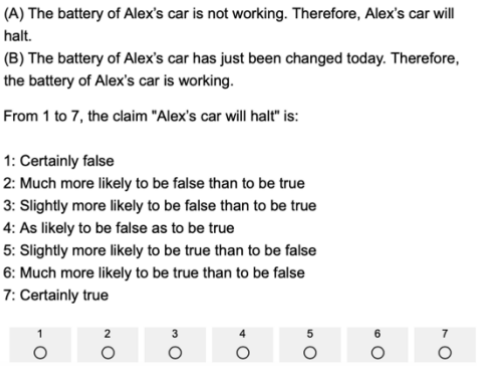

Figure 5b: Defeated stage of set 1

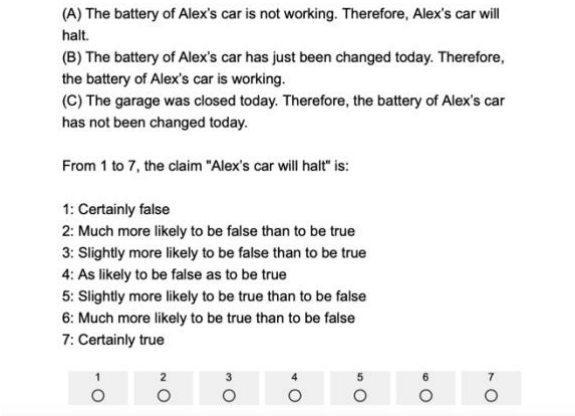

Figure 5c: Reinstated stage of set 1

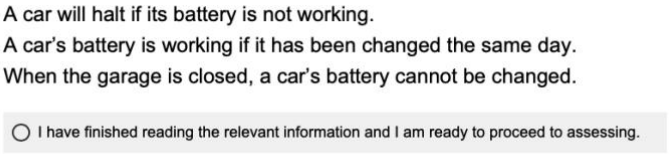

Figure 6: The relevant theory of set 1

Procedure

The participants were provided with a link to their corresponding survey, which they opened on their own device, and were instructed to participate while in a calm environment, to avoid distraction. As previously mentioned, in the beginning participants answered personal questions and questions about their language capacity (which can be found in Appendix A). Then, participants were presented with a brief explanation of what constitutes an argument and an argument's conclusion. After that, specific instructions concerning the procedure were introduced, which concluded the introductory part of the survey.

Participants, then, proceeded to the actual experiment, where they were presented with arguments A, B, and C, and were asked to evaluate the conclusion of argument A. As already mentioned, the way of said presentation varied and each participant was presented with one of the three different manners of presentation, depending on the group to which they belonged. The participant of each group was asked to evaluate four sets of their corresponding case, in a randomised order. The arguments were presented in the following three different ways.

1. The evaluation process consists of three stages. The arguments of Appendix B1 are introduced one by one (i.e., in three different stages), following the order of the attack relations (i.e., the order shown in the graph of Figure 1), ending with the unattacked argument. Participants of group1 are first presented with the base stage, where only A is presented (Figure 5a), then the defeated stage, where A and B are presented (Figure 5b), and, lastly, they are presented with the reinstated stage, where all A, B, and C are presented (Figure 5c). In each stage, participants evaluate the conclusion of argument A.
2. The evaluation process consists of one stage. The arguments of Appendix B1 are all introduced at once, meaning the participants are provided with a text, containing all three arguments, where they are again denoted as arguments A, B, and C. After the participants read the whole text, they evaluate the conclusion of argument A. Essentially, group2 is only presented with the reinstated stage (Figure 5c).
3. The evaluation process consists of four stages. First, the participants are presented with the relevant theory—i.e., with the generalized forms of all the arguments they encounter during the evaluation (Figure 6)—which can be found in Appendix B2. Then the corresponding arguments of Appendix B1 are introduced in the same way that arguments were presented to the participants of the first group (i.e., sequentially). Thus, the participants of group3 are first presented with all possible scenarios abstractly and are then asked to evaluate the conclusion of argument A, in the same way participants of the first group are.

Measurements

The aforementioned hypotheses were tested with manner of argument presentation as the independent variable. In order to test the effect of presentation manner, confidence in the conclusion was used as the dependent variable and one metric was used, i.e., the participants' ratings regarding the acceptability of the conclusion of argument A. The ratings were given on a 7–point scale ranging from *Certainly false* to *Certainly true* (the same that was used by Rahwan et al., 2010). Since responses on individual trials might be influenced by various aspects, I averaged across the four trials of each participant for each case—i.e., for each participant, their average rating over the four sets of arguments was calculated—in an effort to compensate for any noise in the measurement.

In order to test the first hypothesis, the ratings of the first group for the conclusion of argument A in the base stage were compared against the ratings in the reinstated stage. Moreover, the ratings of the first group (i.e., the ratings that are attributed to the conclusion of argument A in the reinstated stage) were compared against the ratings given by the second group for the conclusion argument A, so as to test my second hypothesis, and against the ratings of the third group for the conclusion of argument A in the reinstated stage, for my third hypothesis. Lastly, for my fourth hypothesis, the ratings of the first group in the base stage were compared against the corresponding ratings of the third group.

Results

The first step of the statistical analysis was clearing the data. No participants were found inadequate according to the language standards, but some surveys were incomplete, thus I cleared the data by disregarding participants that had not completed the survey (i.e., I disregarded the data of participants who had dropped the survey at some intermediate stage); 27 participants' data were disregarded from group1 (21% of the size of group1), 31 from group2 (26% of the size of group2), and 20 from group3 (15% of the size of group3). After clearing the data, I proceeded to calculating one average score value per participant per condition. The statistical softwares JASP and R were used for my statistical analysis.

For my first hypothesis, I examined whether my results replicate Rahwan et al.'s (2010) results, i.e., whether ratings given at the base stage of the first group are higher than the corresponding ratings at reinstated stage. To this end, I conducted a comparison within the ratings of the participants of group1. I compared their ratings given at the first stage (when only A is presented) to the corresponding ratings at the third stage (when all three arguments have been presented). A paired t-test found a significant effect of pattern on ratings, $t(102)$ =

4.636, $p < .001$. Ratings were higher in the base stage ($M = 5.61$, $SD = 0.99$, 95% CI = [5.51, 5.71]) compared to the reinstated ($M = 5.21$, $SD = 0.96$, 95% CI = [5.12, 5.30]). Thus, the null hypothesis is rejected and, so, my first hypothesis is confirmed, which means that the results of Rahwan et al. (2010) are successfully replicated. A visualization of said comparison between the ratings in both stages is illustrated in the boxplot below (Figure 7). The conclusion of argument A in the defeated stage was much lower ($M = 3.76$, $SD = 1.16$), as expected.

Moreover, the ratings in the base stage are strongly, positively correlated with the ratings in the reinstated stage—i.e., when the former ratings are higher, so are the latter. Therefore, even though the reinstated ratings are significantly lower than the base ones, the two are related, indicating internal consistency (i.e., that the participants' responses are consistent across items) and, thus, measurement reliability. The aforementioned correlation was confirmed by a Pearson's correlation test ($r = 0.59$, $p < .001$) and is depicted in the scatter plot below (Figure 8). Moreover, on this plot one can see that the majority of participants gave mostly high ratings in both stages, as the majority of points is on the top right part.

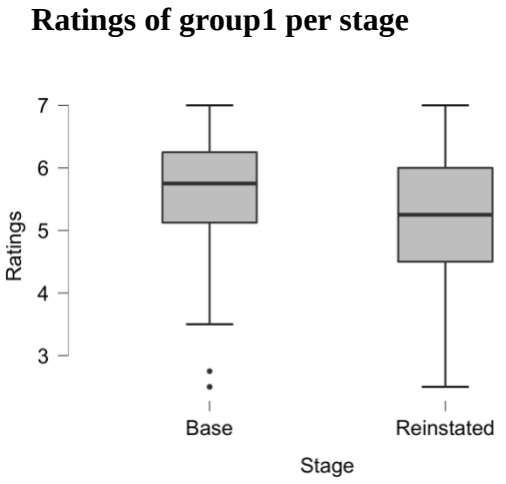

Figure 7: Comparison between first group participants' ratings (y axis) for the two stages; the base stage is depicted on the left, while on the right stands the reinstated stage. The bold horizontal line segment within the boxes represents the mean of the ratings for each case/stage.

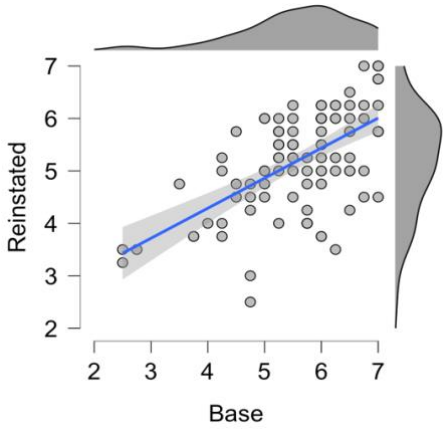

Figure 8: Relation between the ratings of each participant for base stage (abscissa of each point) and for the reinstated stage (ordinate of each point) of group1. The points represent the participants of the first group. On top of the plot stands the ratings' density graph of the base stage and on the right of the plot the one of the reinstated.

Regarding my second hypothesis, I compared the ratings of group2 to the ones of group1 (the ones mentioned as 'reinstated' ratings of group1 until this point, but, for my second hypothesis testing, I will refer to them merely as 'group1'), in order to examine the effect of the manner of argument presentation on the ratings given by the participants. To this end, an

independent *Welch* t-test was performed, since the two samples have unequal sample sizes. The t-test found no significant effect of the presentation manner on rating, $t(196.56) = -0.683$, $p = .496$. Thus, the second hypothesis cannot be confirmed, as there is not enough evidence to reject the null hypothesis. A visualization of said comparison between the ratings of the two groups is illustrated in the graphs below (Figures 9 & 10).

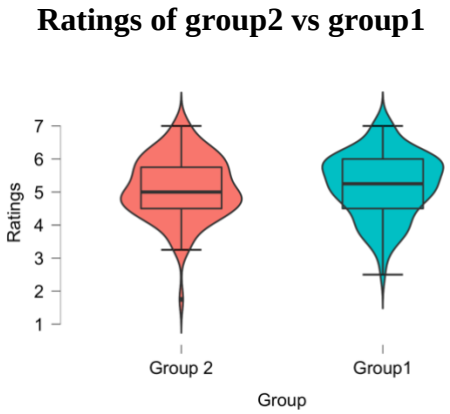

Figure 9: Combination of boxplot and violin plot for the ratings of both groups, where participants' ratings (y axis) are compared; group2 is presented on the left and group1 is on the right. The bold horizontal line segment within the boxes represents the mean of the ratings for each case/group. The violin plot illustrates the distribution of the ratings (the probability density of the data at different values).

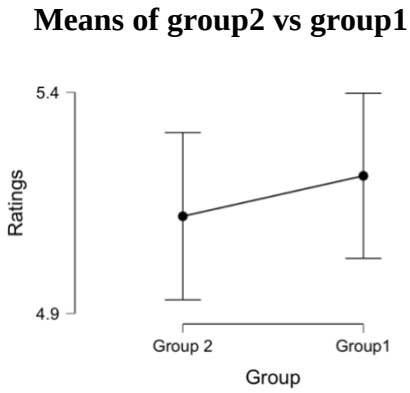

Figure 10: Close up comparison between the means of ratings of both groups. The ordinate of the dots on the whiskers represents said means.

Regarding my third hypothesis, I compared the ratings of group3 in the reinstated stage to the ones of group1 in the reinstated stage. To this end, an independent *Welch* t-test was performed, since the two samples have unequal sample sizes. The t-test found a significant effect of manner of presentation on ratings, $t(207.01) = 2.516$, $p = .013$. Ratings were higher in the third group ($M = 5.53$, $SD = 0.89$, 95% CI = [5.36, 5.70]) compared to the first ($M = 5.21$, $SD = 0.96$, 95% CI = [5.12, 5.30]). Thus, the null hypothesis is rejected and, therefore, my third hypothesis is confirmed. A visualization of said comparison between the ratings of the reinstated stage of the two groups is illustrated in the graphs below (Figures 11 & 12).

For my fourth hypothesis, I compared the ratings of group1 to the ones of group3 in the base stage. To this end, an independent *Welch* t-test was performed, since, evidently, the two samples have again unequal sample sizes. The t-test found again a significant effect of manner of presentation on ratings, $t(201.56) = -3.638$, $p < .001$, but not in favor of my hypothesis. Ratings were higher in the third group ($M = 6.07$, $SD = 0.85$, 95% CI = [5.91,6.23]) compared to the first ($M = 5.61$, $SD = 0.99$, 95% CI = [5.51, 5.71]). Thus, my fourth hypothesis is rejected. From these results, we can conclude that the introduction of the generalized form of all set's

arguments before assessing the conclusion of argument A has an impact on both stages (i.e., the ratings of both base and reinstated stage were significantly different between group1 and group3). A visualization of said comparison between the ratings in the base stage of the two groups is illustrated in the graphs below (Figures 13 & 14).

Reinstated Ratings of group1 vs group3

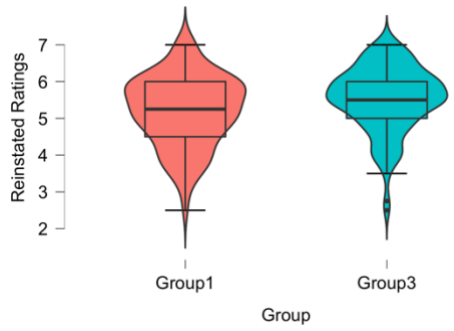

Figure 11: Combination of boxplot and violin plot for the ratings of both groups corresponding to their reinstated stages, where participants' ratings (y axis) are compared; group1 is depicted on the left and group3 is on the right. The bold horizontal line segment within the boxes represents the mean of the ratings for each case. The violin plot illustrates the distribution of the ratings (the probability density of the data at different values).

Reinstated stage Means: group1 vs group3

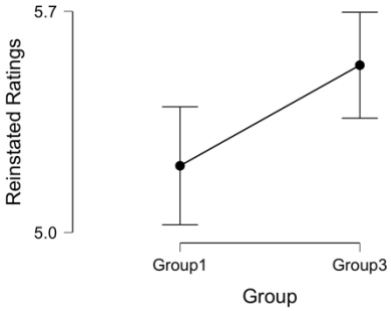

Figure 12: Close up comparison between the means of ratings for the reinstated stage of both groups. The ordinate of the dots on the whiskers represents said means.

Base Ratings of Group1 vs Group3

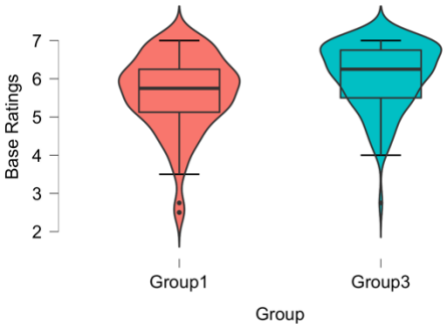

Figure 13: Combination of boxplot and violin plot for the ratings corresponding to the base stage of both groups, where participants' ratings (y axis) are compared; group1 is depicted on the left and group3 is on the right. The bold horizontal line segment within the boxes represents the mean of the ratings for each case. The violin plot illustrates the distribution of the ratings (the probability density of the data at different values).

Base stage Means: group1 vs group3

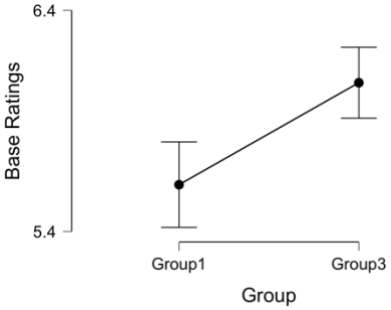

Figure 14: Close up comparison between the means of ratings for the base stage of both groups. The ordinate of the dots on the whiskers represents said means.

On account of the rejection of my fourth hypothesis, contrasted with the confirmation of the third, I conducted some further analysis for the third group. In particular, I performed the same comparison as the one in the first group; a comparison within the ratings of the participants of group3. I compared their ratings given at the base stage with the corresponding ratings at the reinstated stage, in order to check whether reinstatement was imperfect in this case as well. A paired t-test found a significant effect of pattern on ratings, $t(109) = 5.727$, $p < .001$. Ratings were higher in the base stage ($M = 6.07$, $SD = 0.85$, 95% CI = [5.91,6.23]) compared to the reinstated ($M = 5.53$, $SD = 0.89$, 95% CI = [5.36, 5.70]), like in the first group. Thus, even though the participants were introduced to the whole relevant theory beforehand, their ratings in the reinstated stage are still significantly lower, meaning the reinstatement is imperfect again. A visualization of said comparison between the ratings in both stages is illustrated in the boxplot below (Figure 15). The conclusion of argument A in the defeated stage was again much lower, as expected ($M = 3.94$, $SD = 1.27$).

Moreover, the ratings of the base stage and reinstated stage are again positively correlated—i.e., when the former ratings are higher, so are the latter—although at a lower degree compared to group1. Thus, the results indicate internal consistency and measure reliability. The aforementioned, moderate correlation was confirmed by a Pearson's correlation test ($r = 0.35$, $p < .001$) and is depicted in the scatter plot below (Figure 16). Moreover, on this plot one can see that the majority of participants gave high ratings in both stages, as the majority of points is once again on the top right part of the plot.

Ratings of group3 per stage

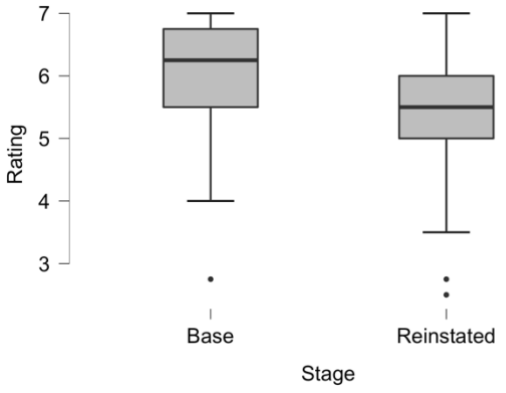

Figure 15: Comparison between third group participants' ratings (y axis) for the two stages; on the left the base stage is depicted, whereas on the right stands the reinstated. The bold horizontal line segment within the boxes represents the mean of the ratings for each case.

Base vs Reinstated stage for group3

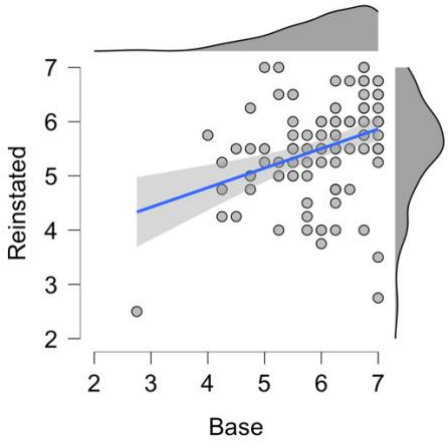

Figure 16: Relation between the ratings of each participant for base stage (abscissa of each point) and for the reinstated stage (ordinate of each point) of group3. The points represent the participants of the third group. On top of the plot stands the ratings' density graph of the base stage and on the right of the plot the one of the reinstated.

Lastly, for a more complete understanding of participants' behavior, I compared the ratings of group3 (previously mentioned as 'reinstated' ratings for group3, but for this comparison I will refer to them merely as 'group3') to the ones of group2, in order to examine whether there is a difference between first introducing an abstract form of all the arguments, and then presenting the arguments sequentially, and presenting all said arguments at once. To this end, an independent *Welch* t-test was performed, since the two samples have unequal sample sizes. The t-test found a significant effect on ratings, $t(197.16) = 3.218$, $p = .002$. Ratings were higher for the third group ($M = 5.53$, $SD = 0.89$, 95% CI = [5.36, 5.70]) compared to the second ($M = 5.12$, $SD = 0.93$, 95% CI = [4.93, 5.31]). A visualization of said comparison between the ratings of the two groups is illustrated in the graphs below (Figures 17 & 18).

Ratings of group2 vs group3

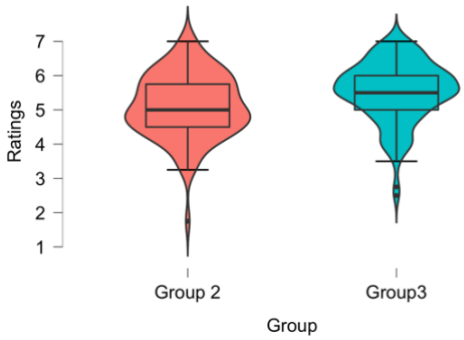

Figure 17: Combination of boxplot and violin plot for the ratings of both groups, where participants' ratings (y axis) are compared; group2 is depicted on the left and group3 is on the right. The bold horizontal line segment within the boxes represents the mean of the ratings for each case. The violin plot illustrates the distribution of the ratings (the probability density of the data at different values).

Means of group2 vs group3

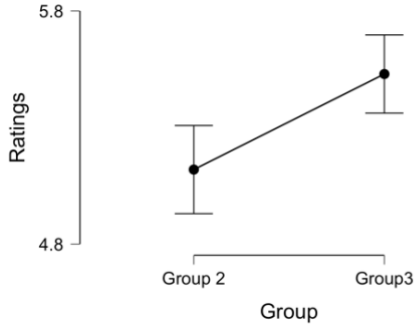

Figure 18: Close up comparison between the means of ratings of both groups. The ordinate of the dots on the whiskers represents said means.

To gain further insight into the data, further analysis was conducted without averaging over the four sets for each participant, in order to examine the sets of arguments individually. Here, again, as 'group1' and 'group3' I consider the reinstated stage of each one of them. For all three groups, the highest rated set was the second set (group1: $M = 5.92$, $SD = 1.18$, group2: $M = 5.70$, $SD = 1.72$, group3: $M = 6.08$, $SD = 1.66$). For the first group, the lowest rated set was the third ($M = 4.25$, $SD = 1.53$), whereas for the second and third group, it was the seventh (group2: $M = 4.28$, $SD = 1.60$, group3: $M = 4.96$, $SD = 1.88$). The content of each set can be found in Appendix B1. It is worth noting that the fact that the eighth set (i.e., the one about animal's rights) was the one that involved world knowledge the most, and perhaps consisted

of the most emotionally charged content, does not appear to have influenced the ratings of the participants (group1: $M = 5.48$, $SD = 1.58$, group2: $M = 5.13$, $SD = 1.56$, group3: $M = 5.66$, $SD = 1.27$), indicating a good level of impartiality from the part of the participants.

Lastly, to make a general comparison: group3 (i.e., its reinstated stage) had overall the highest ratings for the conclusion of argument A ($M = 5.53$, $SD = 0.89$)—being closer to the ratings of the base stage of group1 ($M = 5.61$, $SD = 0.99$) than to its (corresponding) reinstated stage—and group2 had the lowest ($M = 5.12$, $SD = 0.93$). The first group (i.e., its reinstated stage) was in the middle of the two ($M = 5.21$, $SD = 0.96$)—with its ratings being closer to the second group than the third. The low ratings of the second group, along with the fact that group2 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 case was more challenging to the subjects.

General Discussion

Following Rahwan et al. (2010) and their descriptive-experimental approach, this study purported to (1) replicate their findings and (2) investigate some of the criticisms of Prakken & de Winter (2018) regarding said findings constituting supporting evidence for graded semantics (such as semantics proposed by Amgoud & Ben-Naim 2013, Amgoud & Ben-Naim 2016, Grossi & Modgil, 2015, Grossi & Modgil, 2019). To this end, the effect of the manner of argument presentation on the confidence in an argument's conclusion was examined. Four hypotheses were formulated, two of which are confirmed.

Specifically, this research focuses on the matters of (1) suspension of disbelief (and its disruption) and Prakken & de Winter's (2018) variation and (2) graded acceptability. In order to discuss whether imperfect reinstatement supports the latter, this study investigated the former as potential explanations. In this section, these, along with additional explanations—such as directionality and order—are discussed, as well as additional aspects that need to be considered in order to safely conclude that this study's findings, along with the ones of Rahwan et al. (2010), are in support of graded acceptability.

To begin with, the first hypothesis of the current study is confirmed, as the results of Rahwan et al. (2010) were successfully replicated. Such confirmations bear great significance, as replicability is one of the cornerstones of scientific method and social sciences are currently facing a replication crisis. Every successful replication restores the confidence in experimental processes. Regarding the Prakken & de Winter (2018) variation of suspension of disbelief, such explanation is not endorsed by the results of the present study. As Prakken & de Winter (2018)

suggested, if it were true, the initial introduction of the relevant theory should have led to the ratings for the conclusion of argument A in the base stage of group3 being significantly lower than those of group1 (which was hypothesized in my fourth hypothesis). However, my fourth 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 results are even more surprising, at first glance, if we take into consideration that the second hypothesis is also rejected and the ratings of group1 (in the reinstated stage) and group2 were not found significantly different, even though in group2 the participants are again familiar with all possible scenarios (attacker included) before evaluating the conclusion of argument A. Thus, being presented with just the reinstated stage (as was group2) had no significant difference to being introduced gradually to the arguments. Moreover, one would have also expected ratings of group2 and group3 (base and reinstated stage) not to be significantly different, but the results indicate otherwise.

Regarding my third hypothesis, that the ratings in the reinstated stage of group3 are higher than the corresponding ratings of group1, it is confirmed. Therefore, what we can observe is that the introduction of the whole theory beforehand resulted in an overall increased confidence in the conclusion of argument A (i.e., the ratings of both the base and the reinstated stage were higher in group3 than in group1). What is puzzling, thus, is the confirmation of the third hypothesis in contrast with the rejection of the fourth, as what we expected was that the introduction of the theory would have opposite results on the base and reinstated stage. Lastly, the difference between the two stages in the third group was also found significantly different, i.e., the ratings of the reinstated stage were significantly lower than those of the base stage, rendering, thus, the recovery of reinstatement incomplete in this case as well.

Regarding Prakken & de Winter's (2018) variation of the disruption of the suspension of disbelief, despite the confirmation of my third hypothesis—i.e., that that the introduction of the relevant theory results in increased confidence levels in the reinstated stage—two reasons indicate that their explanation cannot be considered an accurate factor of imperfect reinstatement: (1) the ratings for the conclusion of argument A in the base stage of group3 are significantly lower than those of group1 in the base stage; (2) the recovery of reinstatement in the third group is also not complete (same as in the first group). One would expect that if the introduction of an attacker makes participants think of other possible objections (that are not directly addressed in the text), introducing the attacker at the beginning (as it happened in group3), would lower the ratings of the base stage, while raising (or at least leaving the same)

the ratings of the reinstated stage, resulting in the same (i.e., not significantly different) confidence levels in the two stages. But, as observed, results showed a significant difference between the two. Since reinstatement is not complete in this case as well, Prakken & de Winter's (2018) suggestion cannot constitute an explanation of imperfect reinstatement.

Regarding the reasons that the introduction of the corresponding theory results in an increase of the ratings' level in both stages, one explanation 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, there is no reason/evidence not to believe it. On the contrary, though, the introduction of the theory, and 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 realized, 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 group2 (i.e., the version of manner of presentation that corresponded to group2) being more challenging. As mentioned in Cramer & Guillaume (2019), 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). Moreover, they could constitute the results of a confounder, as will be more thoroughly explained in the next section (see Limitations).

Regarding the suggested explanation of Rahwan et al. (2010), in terms of suspension of disbelief, the results of this study cast some doubt on it. Rahwan et al. (2010) 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, I hypothesized that if introduced with all three arguments at the same time, participants would apply their suspension of disbelief to all the (initially) presented arguments. As my second hypothesis 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 (2010) explanation regarding the disruption of suspension of disbelief cannot be validated.

What is more, the findings of this study, along with the ones of Rahwan et al. (2010), are in line with previous studies regarding graded acceptability (such as Amgoud & Ben-Naim 2013, Amgoud & Ben-Naim 2016, Grossi & Modgil, 2015, Grossi & Modgil, 2019). Although,

theoretically, one could speculate that an unattacked argument is less preferred to a reinstated one, since the latter has successfully 'passed a test' posed by its counterargument, Rahwan et al.'s (2010) findings do not confirm this. Prakken & de Winter (2018) also suggested the assumption that, more generally, the more tests an argument manages to survive, the better (as, for instance, in the case of scientific theories); an assumption they base on the theory of Cohen (1977) of Baconian probability. However, again, this study's findings, along with Rahwan et al. (2010) do not confirm this assumption.

However, before being able to safely conclude that said findings support graded semantics, we should first examine alternative explanations. The most important being one that was independently made by Prakken & de Winter (2018) and Cramer & Guillaume (2018), by making use of a structured account to review this problem: *directionality*. In particular (as mentioned in the Related Work section), Rahwan et al. (2010) use undermings as the attack type between each set's arguments, but they assume an asymmetric attack between them. However, such undermings in ASPIC+ can give rise to bidirectional/symmetrical attacks, since in ASPIC+ a premise attack induces a reverse rebuttal, as long as the attacked top rule is defeasible. Although there are different frameworks (such as ABA), where this relation is seen as a unidirectional attack, Cramer & Guillaume (2018) have empirically confirmed that, in the case of (simple) undermining, people judge the attack as bidirectional (as suggested by ASPIC+), rendering Rahwan et al.'s (2010) interpretation problematic. Of course, here we could discuss whether this empirical evaluation is relevant, but this is a point I will be generally addressing later. However, since the study of Rahwan et al. (2010) is an empirical study, which investigates people's behaviours and judgments, the way people interpret the (only) attack used in the experiment is highly relevant.

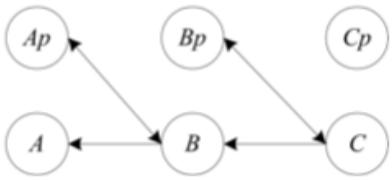

Figure 19:
Graph from Prakken & de Winter (2018) that represents
a structured account of the argument sets used in this experiment.
A_p, B_p, C_p are subarguments of A, B, C, consisting of their premise.
A_p and B directly attack each other and B_p and C directly attack each other.

As has been illustrated by Prakken & de Winter (2018), if we reconstruct the arguments in ASPIC+ (by regarding all premises as ordinary and all inference rules as defeasible), then

the sets that were used in Rahwan et al. (2010) correspond to the framework depicted in Figure 19. A_p, B_p, and C_p are the subarguments of A, B, and C, respectively, and they consist in the arguments' premise. Since B directly attacks A_p, B undermines A on A_p, while A_p directly attacks (rebuts) B, in a symmetric attack. Similarly, C directly attacks B_p, undermining B on B_p, while B_p directly attacks (rebuts) C. The difference between the framework depicted in Figure 1 and the one in Figure 19 is that, although A is sceptically justified in the former, that is not the case in the latter; A is in this case credulously justified (or defensible). Prakken & de Winter (2018) also emphasized that this is a perfect example of the problem of "directly formalising natural-language examples as *AF*s may result in ad-hoc modellings (or in this case in a modelling that is not the only possible one)".

Lastly, what I want to propose is that imperfect reinstatement could be a result of *order*. Specifically, 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 & Hunter (2018): "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 this study and in Rahwan et al. (2010), the arguments were always presented in the same order. Even in group2, where all the arguments were presented together, argument A is always first (on top). We cannot, thus, rule out the possibility that the order of arguments also plays a role in participants' confidence. What is more, it is even conceivable that the order of arguments' introduction has an effect on the way participants perceive the directionality of attacks as well.

Regarding graded acceptability, as previously mentioned, in the current argumentation discourse the findings of Rahwan et al. (2010) have been considered as supporting evidence of graded semantics, as said findings seem to indicate that people follow the central principle that "all else being equal, having fewer attackers is 'better' than having more" (Grossi & Modgil, 2015; Grossi & Modgil, 2019)—which is in accord with one of the main features of the semantics proposed by Amgoud & Ben-Naim (2013; 2019) that "the number of attackers has a great impact on the acceptability of an argument". However, I believe that, despite these indications, one should first examine and rule out the remaining, aforementioned, alternative explanations before interpreting empirically imperfect reinstatement this way.

Nevertheless, let us now examine this claim. As we have seen, Cramer & Guillaume (2018) have also concluded that humans agree with the predictions motivated by ASPIC+ to a varying degree and this could serve as an indication that ASPIC+ is not fine-grained enough. Of course, this could be interpreted as the result of the interference of participants' world

knowledge or confusion. However, potentially, there is something to it and acceptability semantics provides this more fine-grained level. The results of the present study fit with the graded semantics that assign higher ranking to unattacked arguments than to defended arguments, since (1) Rahwan et al.'s (2010) results were replicated, (2) the ratings of group2 were not significantly different from the ones of the reinstated stage of group1, and (3) reinstatement in group3 was incomplete as well.

However, it is important here to make two important distinctions between these empirical findings of imperfect reinstatement and the, more general claim, that fewer attackers make an argument more justified. Both this study and Rahwan et al. (2010) examined a special case, where arguments A and B were compared in the case that AF_1: A and AF_2: D → C → B, while argument A equals argument B. However, as pointed out by Prakken & de Winter (2018), the usual current gradual semantics entail a wider principle that A in AF_1 is more acceptable than B in AF_2, even when A and B refer to different arguments. This gap also appears in the examples Grossi & Modgil (2015; 2019) provide for their principles. For instance, according to their semantics, between A and D, in AF_3: C → B → A and AF_4: F, G→ E →D, D is generally more acceptable than A, even if they are different arguments, but they only provide examples of the special case where A = D and B = E and C = F. However, if they are different, it is possible that they are different in nature; for instance, it could be the case that C is not attackable, while F and G are. This illustrates again the danger of abstraction described by Prakken & de Winter (2018) where making implicit assumptions might not hold in general.

Secondly, we need to distinguish between 'an unattacked argument is better to a defended one' and 'the fewer attackers the better'. Although it seems intuitively natural for the two to be related, the present study provides evidence about the former, which is a special case of the latter. In other words, this study only examines the difference between zero attackers and one attacker. However, there is no evidence that this generalizes to one and two attackers, and so on. The assumption that fewer attackers is better should be further investigated empirically, as this principle is central to the current work on gradual acceptability.

Furthermore, if we pay attention to the results that confirm the incompleteness of reinstatement, the findings are not that unexpected if we take a closer look at the way humans evaluate. Looking at the participants' ratings, an attacker never completely demolishes the confidence in the attacked. For instance, if that was the case, the ratings in the attacked stage of both the first and the third group should have been very close to 1; however, they are far higher (group1: M = 3.76, SD = 1.16, group3: M = 3.94, SD = 1.27). As an attack is not perfect either then, it only makes sense that an attack on an attacker (i.e., a reinstatement) is also not perfect. As argument A's attacker does not completely eradicate confidence in the conclusion

of argument A, it comes as no surprise that the attacker's attacker does not completely eradicate confidence in the conclusion of the attacker, thus not completely restoring confidence in the original argument. This might even suggest that every time an argument is attacked and then reinstated, the confidence in it decreases, but at a decreasing pace. However, here it could also be argued that this is a result of people avoiding extreme ratings, as for instance, in the base stage, argument A's conclusion is not that close to 7 (group1: $M = 5.61$, $SD = 0.99$, group3: $M = 6.07$, $SD = 0.85$). Nevertheless, the difference is more acute in the defeated stage, which I believe to be an indication of a not absolute killing by the attacker.

This is again is in line with graded acceptability semantics by Amgoud & Ben-Naim (2013; 2016)—but could also be perceived as a result of the aforementioned directionality issue again. As mentioned in a previous section (Introduction) Amgoud & Ben-Naim (2013) propose a semantics which suggests that an attack weakens its target but does not kill it. The aforementioned findings could stand in support of this. In particular, one of the basic considerations of their semantics is *weakening*: "[a]rguments cannot be killed (however, they can be weakened to an extreme extent). As a consequence, an attack from an argument b to an argument a always decreases the degree of acceptability of a (possibly only by an infinitesimal amount). The greater the acceptability of b, the greater the decrease in the acceptability of a". A special case of this feature can be observed again in both empirical studies discussed.

Hence, could the results of Rahwan et al. (2010), and of this study, be considered as supporting evidence of graded semantics? Surely, they do not constitute evidence against it, but, at the moment, as these findings are also in line with various alternative explanations (that have not yet been ruled out), and without having tested whether they can be generalized, they cannot help us discriminate between the competing theories/explanations, thus they cannot be used as such evidence (at least not yet).

Of course, the final hurdle one has to overcome in order to take the findings of the present study (as well as the findings of Rahwan et al., 2010, or any empirical study for that matter), as evidence that supports graded acceptability semantics, or any semantics, is: why should we think that what people do is rational? As we have seen, plenty of studies—such as Grossi & Modgil (2015; 2019)—refer to such findings as validation of their theories. Is it right, thus, to assume that people follow normative rules when they argue? As mentioned in the Introduction, Pollock (1986) and Jackson (1988) both advocate that people follow rules and have specific ideas about whether to defend a statement and how to argue. What is more, Jackson (1988) argues that descriptive attempts are merely attempts to reconstruct people's own normative ideas.

As has been also discussed in the beginning of this paper (see The need of empirical evaluation), investigating the reasoning processes of humans can ameliorate our formal theories, and graded acceptability is a good example of this. Amgoud & Ben-Naim (2013) claim that the semantics they propose is more well-suited for certain *applications* of reasoning, such as decision-making. It can be well the case that, depending on the domain and its goals, different semantics are the most appropriate. The motivation behind graded semantics includes applicability and, as previously argued by many (such as Modgil 2013, Prakken 2020, etc.), it makes sense then to be in proximity with 'everyday' human reasoning; just like non-monotonic logics were introduced because standard logic was not suitable for the often defeasible commonsense reasoning. This proximity is the reason we need the make the jump 'from *Is* to *Ought*'.

As suggested by van Benthem (2008) and Rahwan et al. (2010), when assessing formal models of reasoning, argumentation, and decision making, cognitive plausibility should play a role, while careful experimental design can help shape formal theories. What we need to decide is in what ways empirical evaluation and validation are most beneficial to argumentation theories. As suggested by Prakken (2020), the goal is to state theories "in terms that are natural to people". An empirical validation can serve as guidance for researchers, giving us insights into semantics, or providing 'solutions' when semantics are in disagreement (Jackson, 1988). The empirical inquiry's findings are not something to be followed to the letter, but they can serve as a compass that helps us navigate, while human reasoning can constitute a place to resort to when conflict between normative theories is generated.

Limitations & Future Research

The biggest limitation of this study, as perhaps of most studies that consist of similar empirical experiments, is the interference of participants' world knowledge. As already mentioned, in reasoning experiments, like the one conducted in this study, it is very hard to control to what degree the participants base their answers on the provided information, as pointed out by Prakken & de Winter (2018), and in line with the findings of multiple studies, such as Evans & Over (2004), Cerutti et al. (2014), Cramer & Guillaume (2018), and Polberg & Hunter (2018).

Although explicitly instructed not to do so in the beginning of the survey, it is possible that the participants were influenced by other (own) beliefs and general background information. World knowledge is a particularly dominant interference as we already saw in Cramer & Guillaume (2018), where participants judged statements based on their implausibility, and in Cerutti et al. (2014), where ratings depended on the domain of the

arguments. Moreover, Evans & Over (2004) found that in reasoning experiments participants will most likely judge the soundness of an argument, even when explicitly asked to judge the validity of their conclusions. Moreover, they found that when reasoning about 'causal' conditionals p → q, prior knowledge influences the perception of both the sufficiency of p for q and of its necessity; "we will not draw the valid MP and MT inferences if we can think of disabling conditions that would prevent p leading to q" (Evans & Over, 2004).

In this study, the content of the arguments that were used were simple sentences and, in their majority, of a neutral subject matter (to avoid subjective views influences). Moreover, the levels of confidence in the eighth set (i.e., the one about animal rights), which is the most emotionally engaging, suggest a good level of impartiality from the participants. Nevertheless, we cannot exclude the possibility that the results are partly influenced by the *content* of the arguments (for instance, perhaps some arguments were perceived as less persuasive or plausible by the participants), rather than their *relations*. Future experiments could include a manipulation check, where a separate group of participants evaluates the arguments independently in order to control for this (for instance, check whether the attacker has a significantly higher confidence level than the argument it attacks and, if so, whether that plays a role in the confidence level of the attacked argument). Furthermore, in this direction, future research could also try to investigate this effect further (as did Cerutti et al., 2014), rather than merely trying to eliminate it—as it provides meaningful insights to 'actual' argumentation—by introducing more complex, or emotionally engaging, or domain-varying argument sets.

Regarding the second group, the fact that it had the lowest ratings, along with the fact that group2 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 case was overall challenging to the subjects. What is more, another confounding factor that could have influenced the ratings of this group (in contrast with the other two) is that participants in group2 were only asked to evaluate one time for each set. Thus, said participants were less 'trained' in evaluating, during the process, compared to the participants of the other groups that performed three evaluations per set, thus acquiring a more 'evaluating mentality'.

Moreover, this study did not test any of the alternative explanations discussed in the previous section, such as directionality or order. In particular, this study did not check how participants interpreted the argument attacks used (i.e., as unidirectional or bidirectional). Future research could again make use of manipulation checks, where a separate group of participants are asked to specify the attack relations between the presented arguments.

Regarding the order that the arguments were presented, this study did not control for it. As mentioned in the previous section, it is possible that order also has an effect on participants' ratings. In this study, argument A was always presented first (always followed by B, then C), thus the temporal element should be investigated further; would a different order of presentation of the arguments influence the results (and how)? For instance, it is possible that results would be different if the attacker, argument B, was presented first. Or it is even conceivable that participants might feel the need to slightly change their answer every time they are presented with new information (which could explain the difference between base and reinstated stage in group3). Moreover, is it also possible that order has an effect on the way participants perceive the aforementioned directionality of attacks. Future research should investigate this temporal element of argument presentation, with regard to confidence in arguments' conclusions, but also the participants' perception of the attacks' directionality.

Besides the order of the arguments and the directionality of the attacks, further research should also investigate different *types* of attack. In Rahwan et al. (2010), as well as in this study, the only attack type that was used was (simple) undermining. The phenomenon of imperfect reinstatement should, thus, be investigated with other types of attack as well (both unidirectional and bidirectional). Future research should, thus, examine whether this phenomenon appears with other kinds of attack and whether the type of attack has an effect on participants' ratings. Of course, optimally, each of these studies should first include the aforementioned manipulation check, so that they can verify how each attack type is perceived.

Lastly, in order to safely conclude that such findings can be used as supporting evidence for graded acceptability semantics, future research should also investigate whether these findings generalize in the wider cases mentioned in the previous section. As previously explained, the examples used in this study are just a special case of the principles of graded semantics. Future research should compare cases where the arguments under comparison (i.e., the arguments whose attackers we vary in number) are not the same argument. Moreover, in order to examine whether "fewer attackers are better than more" stands for cases where neither of the compared arguments is unattacked, future research should develop experiments that compare, at least, having one attacker with having two attackers (and, of course, the more cases one tests, and the higher the number of attackers one tests, the more concrete becomes the theory). What is more, future experiments that make use of a linear structure of multiple attacks (for example argument A being attacked by B, which is being attacked by C, which is being attacked by D, etc.) could also investigate whether "fewer attackers are better than more" still applies, as well as whether every time an argument is attacked and then reinstated, the confidence in it decreases at a decreasing pace.

Conclusion

The formal study of argumentation is a highly important branch of AI and empirical findings concerning imperfect simple reinstatement are currently in the centre of argumentation discourse. This study followed a descriptive-experimental approach to empirically examine cases of simple reinstatement.

Firstly, this study purported to replicate the results of the empirical study of Rahwan et al. (2010) regarding the phenomenon of imperfect reinstatement (i.e., the empirical finding that participants' confidence in the conclusion of an argument does not return to its original levels even after the introduction of a counter-attacker that defends it). The first motivation of this study was, therefore, to try and replicate said findings, because, although replicability is a vital component of the scientific process, it has been a big issue in cognitive science the past years. The reason this study focused on the findings of Rahwan et al. (2010) for simple reinstatement is that they have recently gained great importance, as they have become central in today's argumentation discourse regarding graded acceptability. The second aim of this study was, thus, to investigate whether said findings are rightly being used as supporting evidence of particular features of models of graded argument acceptability and, specifically, whether they can be explained by the assumption that people follow the normative "all else being equal, having fewer attackers is 'better' than having more" (Grossi & Modgil, 2015; Grossi & Modgil, 2019). To this end, this study investigated some of the criticisms of Prakken & de Winter (2018) against these appeals.

More specifically, this study consisted of an experiment which, additionally to replicating Rahwan et al.'s (2010) findings, purported to verify two alternative explanations of them: the disruption of the suspension of disbelief, suggested by Rahwan et al. (2010), and its variation by Prakken & de Winter (2018). To this end, the effect of the manner of argument presentation on the confidence in an argument's conclusion was examined. The results of the experiment successfully replicated the findings of Rahwan et al. (2010) but could not confirm either explanation. In particular, it was found that introducing all arguments at once has no effect on the ratings of participants' confidence, while the introduction of the relevant theory before evaluating the arguments increased the ratings level in both base and reinstated stage of simple reinstatement. Moreover, in this latter case, reinstatement was again imperfect (i.e., recovery was incomplete).

The results of this study are in line with the predictions of graded acceptability semantics (such as Amgoud & Ben-Naim 2013, Amgoud & Ben-Naim 2016, Grossi & Modgil, 2015, Grossi & Modgil, 2019). However, in order for one to consider these findings as

supportive of the aforementioned semantics, and, in particular, of the principle "all else being equal, having fewer attackers is 'better' than having more" (but also "an attack weakens its target but does not kill it"), there are still extra steps that need to be taken.

To begin with, this study only investigated *some* of the criticisms of Prakken & de Winter (2018). Thus, the first thing that needs to be done is to rule out all other potential alternative explanations of imperfect reinstatement (i.e., all other than the one suggesting that humans apply the aforementioned normatives). Other than the two examined in this paper, alternative explanations are: directionality (i.e., that the attack used between the arguments is interpreted as symmetrical by participants, rather than unidirectional) and order (suggesting that the order of argument presentation is a confound that was not controlled for in Rahwan et al., 2010, or in this study, affecting participants' ratings).

Moreover, even if this is achieved, additional research should be conducted (or leaps of faith taken) in order to verify whether these empirical findings can generalize in all the ways graded semantics' theories predict. Firstly, the findings confirm that fewer attackers are better if argument A and B are the same in this case: AF_1: A and AF_2: D → C → B. However, the usual current gradual semantics entail a wider principle that A in AF_1 is more acceptable than B in AF_2, even when A and B refer to different arguments. Secondly, this study confirms the special case that being unattacked is better to being reinstated. It needs to be investigated further whether this can be generalized when more attacker and defenders are added (for instance, comparison between having three attackers to four attackers).

To conclude, the findings of this study, as well as the ones of Rahwan et al. (2010), are in line with graded acceptability semantics, but they only confirm a special case of its principles, while they can also be interpreted as results of different factors, which have not been investigated. Therefore, said findings do not bear discriminatory power against their competing explanations, and, thus, they cannot yet constitute supporting evidence for graded acceptability semantics. Additional research is needed first. Of course, in order to take empirical findings as validating of a normative theory, one should also decide whether empirical validation bears any significance to normative theories. The former reflects how people think, while the latter represents how people should think. Two questions persist: do we need an empirical evaluation, and can we assume that the way people reason is rational? These are two questions that anyone who wants to interpret imperfect/incomplete reinstatement as supporting evidence of graded acceptability should answer positively. Personally, I believe that people do, to an extent, think rationally and that it is important for argumentation theories to take human reasoning into account by means of similar empirical studies.

Bibliography

Amgoud, L., & Ben-Naim, J. (2013). Ranking-based semantics for argumentation frameworks. In *Proceedings of the International Conference on Scalable Uncertainty Management* (pp. 134-147). Springer, Berlin, Heidelberg.

Amgoud, L., & Ben-Naim, J. (2016). Evaluation of arguments from support relations: Axioms and semantics. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (IJCAI) (pp. 900–906). New York.

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

Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., & Toni, F. (2014). Introduction to structured argumentation. *Argument & Computation*. *5*(1), 1-4.

Cayrol, C., & Lagasquie-Schiex, M. C. (2005). Graduality in argumentation. *Journal of Artificial Intelligence Research*, *23*, 245-297.

Cerutti, F., Tintarev, N., & Oren, N. (2014, August). Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *Proceedings of the 21st European Conference on Artificial Intelligence* (pp. 207-212).

Cohen, L. J. (1977). *The Probable and the Provable*. Clarendon Press, Oxford.

Cramer, M., & Guillaume, M. (2018). Directionality of attacks in natural language argumentation. In C.Schon (Ed.), *Proceedings of the fourth Workshop on Bridging the Gap between Human and Automated Reasoning* (Bridging 2018) (pp.40-46). http://ceur-ws.org/Vol-2261/paper7.pdf.

Cramer, M., & Guillaume, M. (2018). Empirical cognitive study on abstract argumentation semantics. In S.Modgil et al. (Ed.), *Computational Models of Argument: Proceedings of COMMA 2018* (pp. 413-424). Amsterdam, IOS Press.

Cramer, M., & Guillaume, M. (2019). Empirical study on human evaluation of complex argumentation frameworks. In *Proceedings of the European Conference on Logics in Artificial Intelligence* (pp. 102-115). Springer, Cham.

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), 321-357.

Dung, P. M., Kowalski, R. A., & Toni, F. (2006). Dialectic proof procedures for assumption-based, admissible argumentation. *Artificial Intelligence*. *170*(2), 114-159.

Eemeren, F. H. van, Grootendorst, R., Jacobs, C. S., & Jackson, S. A. (2002). *Reconstructing Argumentative Discourse*. The University of Alabama Press, Alabama.

Evans, J. S. B., & Over, D. E. (2004). *If: Supposition, pragmatics, and dual processes.* Oxford University Press, USA.

Grossi, D., & Modgil, S. (2015). On the graded acceptability of arguments. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence* (IJCAI) (pp. 868-874). AAAI Press.

Grossi, D., & Modgil, S. (2019). On the graded acceptability of arguments in abstract and instantiated argumentation. *Artificial Intelligence*. *275*, 138-173.

Jackson, S. (1988). What can argumentative practice tell us about argumentation norms. In *Norms in argumentation. Proceedings of the Conference on Norms* (pp. 113-122). De Gruyter Mouton.

Kahneman, D. (2012). A proposal to deal with questions about priming effects. An open letter to the scientific community: http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf.

Modgil, S. & Prakken, H. (2012), Resolutions in structured argumentation. In B.Verheij et al. (Ed.), *Computational Models of Argument. Proceedings of COMMA 2012* (pp. 310–321). Amsterdam, IOS Press.

Modgil, S., Toni, F., Bex, F., Bratko, I., Chesnevar, C. I., Dvořák, W., Falappa, M. A., Fan, X., Gaggl S. A., García, A. J., González , M. P., Gordon, T. F., Leite, J., Možina, M., Reed, C., Simari, G. R., Szeider, S., Torroni, P., & Woltran, S. (2013). The added value of argumentation. In Ossowski S. (Ed.), *Agreement technologies* (pp. 357-403). Springer, Dordrecht.

Polberg, S., & Hunter, A. (2018). Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *International Journal of Approximate Reasoning*. *93*, 487-543.

Pollock, J. L. (1986). *Contemporary Theories of Knowledge.* Rowman & Littlefield, NY.

Prakken, H. (2020). On validating theories of abstract argumentation frameworks: the case of bipolar argumentation frameworks. In F.Grasso et al. (Ed.), *Proceedings of the 20th Workshop on Computational Models of Natural Argument* (CMNA 2020) (pp.21-30). http://ceur-ws.org/Vol-2669/paper3.pdf.

Prakken, H., & Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities. *Journal of Applied non-Classical Logics*. *7*(1-2), 25-75.

Prakken, H. & de Winter M. (2018). Abstraction in argumentation: necessary but dangerous. In S.Modgil et al. (Ed.), *Computational Models of Argument: Proceedings of COMMA 2018* (pp. 85-96). Amsterdam, IOS Press.

Rahwan, I., Madakkatel, M. I., Bonnefon, J. F., Awan, R. N., & Abdallah, S. (2010). Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science. 34*(8), 1483-1502.

Rahwan, I., & Simari, G. R. (Eds.). (2009). *Argumentation in artificial intelligence* (Vol. 47). Heidelberg: Springer.

Veltman, F. (1987). *Logics for Conditionals* (Doctoral dissertation, University of Amsterdam, Amsterdam).

Appendices

Appendix A: Language Questionnaire

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:

    a. Which one?
    b. What was the result (what is your level according to it)?

Appendix B: Materials used

B1: The argument sets

Sets 1-6 are taken from Rahwan et al. (2010).

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.

B2: Corresponding generalisations of argument sets of B1

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 have 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 break 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.