Journal of Applied Logic 7 (2009) 403–420

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

Two approaches to the problems of self-attacking arguments and general odd-length cycles of attack

Gustavo A. Bodanza *, Fernando A. Tohmé

Logic and Philosophy of Science Research Center (CILF) and Artificial Intelligence Research and Development Laboratory (LIDIA), Universidad Nacional del Sur, Bahía Blanca, CONICET, Argentina

ARTICLE INFO

Article history:
Received 8 March 2007
Accepted 6 June 2007
Available online 15 June 2008

Keywords:
Argumentation frameworks
Self-attacking arguments
Odd-length cycles of attack
Credulous semantics
Lottery paradox
Game-theory

ABSTRACT

The problems that arise from the presence of self-attacking arguments and odd-length cycles of attack within argumentation frameworks are widely recognized in the literature on defeasible argumentation. This paper introduces two simple semantics to capture different intuitions about what kinds of arguments should become justified in such scenarios. These semantics are modeled upon two extensions of argumentation frameworks, which we call *sustainable* and *tolerant*. Each one is constructed on the common ground of the powerful concept of admissibility introduced by Dung in [P.M. Dung, On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming, and *n*-person games, Artificial Intelligence 77 (1995) 321–357]. The novelty of this approach consists in viewing the admissibility of a subset of arguments as relative to potentially challenging subsets of arguments. Both sustainable and tolerant semantics are more credulous than preferred semantics (i.e. they justify at least the same arguments, and possibly more). Given certain sufficient conditions they coincide among them as well as with other semantics introduced by Dung.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Argument systems are aimed to capture the intuition that actual reasoning proceeds by means of the construction and confrontation of defeasible arguments. A tentative conclusion may be revised if its supporting arguments are attacked by others, leading to new tentative conclusions, and so on. Several formalisms have been proposed to model this idea, each focusing on different details, spanning from the internal structure of arguments to the characterization of the warranted or justified arguments.

Despite their differences, the development of argument systems has led to some common paradoxes or, at least, problematic issues. Some of them where inherited from early approaches to non-monotonic reasoning, such as the well-known *Nixon's Diamond* where two arguments attack each other. To solve this problem two intuitions have been invoked, one that postulates that none of the conflicting arguments should be warranted—the "skeptical" view—while the other states that it is right to grant equal warrant to both of them, so that the user of the system can freely select either one—the "credulous" view.

Another kind of problems arose from the process of argumentation itself, namely the occurrence of odd-length cycles of attack. In the most simple case, one argument attacks itself. Pollock [16] studied a particular instance derived from the *lottery paradox*, introduced by Kyburg [12] as a problem for reasoning with probabilities. Imagine a fair lottery with 1,000,000 tickets, so that each ticket has one in a million chances of winning. Given a particular ticket, we can tentatively

* Corresponding author.
E-mail addresses: ccbodanz@criba.edu.ar (G.A. Bodanza), ftohme@criba.edu.ar (F.A. Tohmé).

1570-8683/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jal.2007.06.012

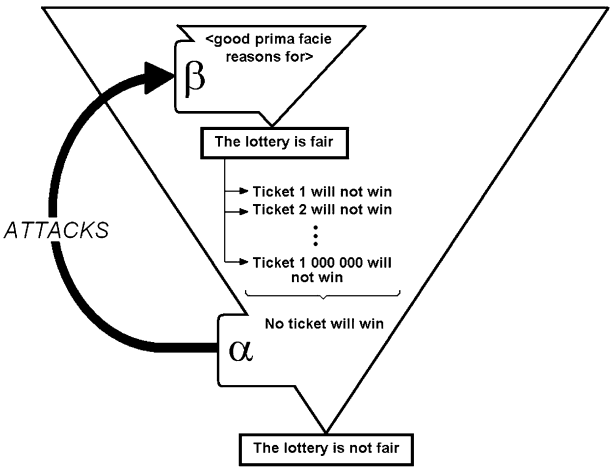

Fig. 1. The "lottery paradox" paradox.

conclude that it will not win, since its probability is very low. But the same can be concluded about *every* ticket, hence we can conclude that none of them will win. But given that the lottery is fair, one ticket *must* win. Upon this contradiction Pollock claimed that a further contradiction can be drawn, namely that since none of the tickets will win, the lottery cannot be fair. That is, the argument of the lottery paradox, based upon the premise that the lottery is fair attacks itself, since one of its consequences contradicts its premise. This is what Pollock called the '*lottery paradox* paradox' (see Fig. 1).

Self-attacking arguments are cycles of attack of length one. For odd-length cycles, Dung [10] indicates that arguments in such cycles should be deemed *controversial* since they both attack and defend, directly or indirectly, some arguments (in particular, themselves). Assume that A attacks B, B attacks C and C attacks A. Since A attacks C's attacker, A defends (i.e. attacks the attacker of) C. Hence, A defends its own attacker, so A indirectly attacks itself. The same happens in any cycle of odd length. This analysis led Dung (and most of the researchers in the field) to try to prevent the occurrence of such arguments within any *extension*, i.e. set of warranted arguments.

In this paper we propose two semantics for dealing with the problem of self-attacking arguments. We define them upon the methodological artifact of seeing an argumentation process as a strategic interaction between two agents. We do not attach ontological weight to this assumption, since the agents may be the two sides that a single reasoner may take while pondering which arguments should be warranted.

Our first semantics takes up from a wide consensus among researchers in the argument systems' community, who seek to solve the problem of self-attacking arguments by rejecting any justification for arguments in any odd-length cycles of attack. We define a new kind of extension called "sustainable", capturing this lead. The second semantics solves as well the problem of self-attacking arguments, but allows the justification of arguments belonging to longer odd-length cycles of attack. Although this runs against the consensus, some authors are also beginning to pursue similar aims [3,4]. The rationality of this semantics, captured by extensions called "tolerant", becomes clear when the attack relation among arguments is seen as embodying a transitive preference relation. Then, as we will show, tolerant semantics exhibit the intended behavior of preferred semantics.

The platform on which our semantics are developed is the argumentation frameworks model of Dung. The main idea introduced there is the *admissibility* of sets of arguments. This notion defines the way a set of arguments can be defended from external attacks. We build upon this idea, stating that the defense of a set of arguments should not be defined in absolute terms but relative to other possible challenging sets of arguments. In this way we obtain a notion of *cogency* on which we base both sustainable and tolerant semantics.

The intuition behind is that an acceptable set of arguments is not necessarily one that can be defended against uncoordinated attacks, but one that can be defended against arguments that are firmly supported. Philosophers of science will recognize the similarity of this condition with the one stating that the withdrawal of scientific theories is not due to some contrary evidence, but to the appearance of a better theory that supports that evidence. In this vein, this paper also proposes a game-theoretic counterpart of sustainable and tolerant semantics showing a strategical rationale for adopting them. A strategic argumentation game will be defined as a two-agent game in which each agent has to choose a set of arguments to confront with and defend against the possible choices of the other agent.

The paper is organized as follows. The main features of Dung's system are presented in Section 2. In Section 3 we introduce the two semantics. In Section 4 we show that in well-founded argumentation frameworks (those in which no cycles of attack occur), sustainable, tolerant and preferred semantics are equivalent. In Section 5 we argue for the rationality of tolerant semantics by showing that the hypothesis *attacks* = *conflict* + *preference* establishes also a sufficient condition for the equivalence between this and preferred semantics. The game-theoretic counterpart of the semantics is offered in Section 6, arguing for strategical foundations. In Section 7 we analyze and discuss related work, while the main conclusions are summarized in Section 8. Appendix A provides the proofs of the formal results.

2. Dung's argumentation frameworks

The formal platform on which this work is based is Dung's argumentation frameworks [10]. An *argumentation framework* is a pair $AF = \langle AR, attacks\rangle$, where AR is a set of abstract entities called *arguments* and *attacks* is a binary relation $attacks \subseteq AR \times AR$ intended to denote attacks among arguments. The fundamental question about argument frameworks is which arguments remain *justified* or *warranted* under the attacks among them. A set of justified arguments is called an *extension*, and alternative criteria yield different kinds of extensions. Two basic concepts allow to obtain all the extensions defined by Dung: *acceptability* of arguments, and *admissibility* of a set of arguments. The formal definitions of these notions are summarized as follows:

Definition 1. (See Dung [10].) In any argumentation framework AF an argument σ is said *acceptable* w.r.t. a subset S of arguments of AR, in case that for every argument τ such that τ *attacks* σ, there exists some argument $\rho \in S$ such that ρ *attacks* τ.[1] A set of arguments S is said *admissible* if each $\sigma \in S$ is acceptable w.r.t. S, and is conflict-free, i.e., the attack relation does not hold for any pair of arguments belonging to S. A *preferred extension* is any maximally admissible set of arguments of AF. A *complete extension* of AF is any conflict-free subset of arguments which is a fixed point of $F(\cdot)$, where $F(S) = \{\sigma\colon\ \sigma$ is acceptable w.r.t. $S\}$, while the *grounded extension* is the least (w.r.t. $\subseteq$) complete extension. Moreover, a *stable extension* is a conflict-free set S of arguments which attacks every argument not belonging to S.

Extensions are said to constitute the "semantics" of an argumentation framework, because they capture the notion (weaker than truthfulness) of *justifiability* of arguments. Different kinds of extensions give way to alternative semantics, that is, different intuitions about which arguments can be justified. The usual interpretation is that semantics justifying a single extension correspond to some kind of skeptical arguer. On the other hand, in a semantics that yields multiple extensions, to select their intersection also characterizes a kind of skeptical attitude, while the arbitrary choice of one of the extensions represents a credulous decision. For example, in Dung's system, preferred, complete and stable extensions can be used to model alternative forms of credulity, while the grounded extension and the intersections (each for each notion of credulity) of the credulous extensions can model different conceptions of skepticism.

3. A strategical approach to argumentation

Dung's view of argument justification does not assign any role whatsoever to the utterer/s of the arguments. His framework can be interpreted as the decision-making setting of a single agent, who ponders the arguments in the framework. Another interpretation is to see the arguments as laid out by an agent to defend herself against the arguments presented by another agent. This distinction motivated some researchers to view defeasible argumentation from a dialogical point of view, modeling dialogues between two players (*pro* and *con*) who develop an argumentation strategy by advancing arguments one by one in alternation (e.g. [1] and [24]).

While we take this view of argumentation as arising from the interaction of two agents, the novelty of our approach consists in that we focus on *argumentation strategies* instead of considering the properties of a sequential dialogue. An argumentation strategy can be viewed as the set of arguments that an intentional agent may lay out in response to the arguments of the other agent in a dialogue. As this notion involves all possible combinations of arguments, we consider that any subset of AR in an argumentation framework $AF = \langle AR, attacks\rangle$ can be seen as an argumentation strategy. In this context, our main working hypothesis is that the rationality of choosing an argumentation strategy is relative to the argumentation strategies that the rival may select. What matters here is not whether an argumentation strategy (as a subset of arguments) is "admissible" in the whole framework, but if it is so relative to the sub-frameworks formed by the strategy itself together with the possible strategies of the other agent.

Under this view we can rethink the main problems of argumentation. While usual questions about argumentation ask how *arguments* can be compared and justified, we want to know now how *argumentation strategies* can be compared and justified. Once that question is settled, we extend the justification to individual arguments. The main methodological steps in the characterization of justification are based on the following assumptions:

- argumentation strategies can be compared on the basis of the attack relation among the arguments that they include,
- argumentation strategies can be justified on the basis of their comparison, and
- the justification of an argument obtains from the justification of the argumentation strategies to which it belongs.

The full strategic conception of argumentation will be spelled out in Section 6. Before that, we will introduce our two credulous semantics. Later we will justify them in game-theoretic terms.

[1] In order to simplify the notation we will say that σ *attacks* A (conversely, A *attacks* σ) iff there exists some $\tau \in A$ such that σ *attacks* τ (τ *attacks* σ, respectively).

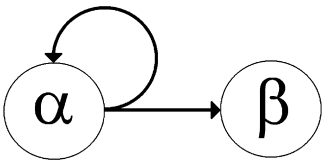

Fig. 2. Argumentation framework of the "lottery paradox" paradox: α attacks its subargument β, hence it attacks itself.

3.1. Admissibility and cogency

The first idea is to establish a comparison among argumentation strategies (just 'strategies' henceforth) using the notion of admissibility. We will consider a strategy A *at least as cogent as* a strategy B if and only if A is conflict-free and all its arguments can be defended against the arguments of B. This amounts to saying that A is an admissible set within the framework formed just by the arguments in $A \cup B$. In order to introduce this definition, we first need an auxiliary notion: a *restriction* of an argumentation framework AF by a subset of arguments $S \subseteq AR$, is the framework $AF\downarrow_S = \langle S, attacks\downarrow_S\rangle$, where $attacks\downarrow_S$ is the restriction of *attacks* to S. Then:

Definition 2. For any pair of strategies $A, B \subseteq AR$, we say that A is *at least as cogent as* B, in symbols '$A \hookrightarrow B$' iff A is admissible in $AF\downarrow_{A\cup B}$.

It is easy to see that:

Proposition 3. *$A \subseteq AR$ is admissible in AF iff for every $B \subseteq AR$, $A \hookrightarrow B$.*

Note that according to this result, it is easy to recover the preferred semantics by requiring a strategy A to be a maximal strategy (for set-theoretic inclusion) such that for every $B \subseteq AR$, $A \hookrightarrow B$.

Admissibility is sufficient for cogency, but cogency is weaker than admissibility: while we understand that A is admissible if it can be defended against any other strategy, A is just cogent if it can be defended against any other strategy which can be defended against A. Formally:

Definition 4. A strategy $A \subseteq AR$ is *cogent* (in AF) iff for every strategy $B \subseteq AR$, if $B \hookrightarrow A$ then $A \hookrightarrow B$.

3.2. Sustainable semantics

Our first semantics arises just asking for a strategy to be a maximally cogent strategy. We call it 'sustainable' suggesting a strategy that holds against any articulated criticism.

Definition 5. A subset $A \subseteq AR$ is a *sustainable extension* of AF iff A is a maximal (w.r.t. $\subseteq$) cogent strategy in AF.

Sustainable extensions avoid the undesirable interference of self-attacking arguments, as shown in the following example:

Example 6. In the argumentation framework $\langle\{\alpha,\beta\},\{(\alpha,\alpha),(\alpha,\beta)\}\rangle$ (see Fig. 2), the set of all possible strategies is $2^{\{\alpha,\beta\}} = \{\emptyset,\{\alpha\},\{\beta\},\{\alpha,\beta\}\}$. Then, $\{\beta\}$ is the only sustainable extension since, besides itself, $\emptyset$ is the only strategy as cogent as $\{\beta\}$, but $\{\beta\}$ is as cogent as $\emptyset$ and is larger. Note that $\{\alpha\}$ is not as cogent as $\{\beta\}$, so $\{\beta\}$ needs not be as cogent as $\{\alpha\}$.[2]

Example 7. In the argumentation framework $AF = \langle\{\alpha,\beta,\gamma\},\{(\alpha,\beta),(\beta,\gamma),(\gamma,\alpha)\}\rangle$, the only sustainable extension is $\emptyset$.

It is easy to see that sustainable extensions yield a more credulous semantics than preferred extensions. Example 6 is a case in which the only preferred extension, $\emptyset$, is a subset of the only sustainable extension, so the latter is clearly more credulous than the preferred extension. We can generalize this fact:

Proposition 8. *If A is a preferred extension, there exists some sustainable extension B such that $A \subseteq B$.*

Given a preferred extension, the sustainable extension to which it belongs may be the preferred extension itself. So, in Example 7, $\emptyset$ is both the only preferred and sustainable extension.

[2] This can be interpreted as an impossibility of shifting the burden of proof from $\{\alpha\}$ to $\{\beta\}$, that is, there is no need of proving the admissibility of $\{\beta\}$ in $AF\downarrow_{A\cup B}$, unless its attacker $\{\alpha\}$ is proved to be admissible in that framework.

3.3. Tolerant semantics

Let us now consider more general cases of odd-length cycles of attack. The behavior we want to capture is the following: while in the framework of Example 6 we would like to justify β, in the framework of Example 7 we should justify just one of the three arguments in the cycle. Later in Sections 5 and 6 we will argue that to do so is rational, but for now we will build further intuition considering interpretations of the framework of Example 7.

The first is due to Pollock [18], in which the three-cycle of attacks is produced by the incompatible statements of three witnesses:

α: "Jones says that Smith is unreliable, hence Smith is unreliable";
β: "Smith says that Robertson is unreliable, hence Robertson is unreliable;
γ: "Robertson says that Jones is unreliable, hence Jones is unreliable".

For practical purposes it could be advisable to accept either one of these arguments, rejecting as excessively skeptical the view that every witness lies.

A similar conclusion can be drawn in a context where three different scientific paradigms are in conflict. We can put this in terms of a scientific discussion that could have taken place at the beginning of the XVIIth century. Assume the arguments are:

α: "There exists evidence supporting Copernicus' theory and his theory is simpler than Ptolemy's, therefore Ptolemy must be wrong";
β: "There exists evidence supporting Ptolemy's theory and his theory is quite in accordance with the Holy Bible, contrary to what happens with Tycho's theory, therefore Tycho must be wrong";
γ: "There exists evidence supporting Tycho's theory which, in accordance with the Scriptures, explains that the Sun revolves around the Earth, therefore Copernicus must be wrong".

Note that α, β and γ provide reasons against the premises of β, γ and α respectively, so we have a cycle of attacks of length three. Even if choosing either one of α, β or γ implies to adhere to a less than certain position, each of these arguments could have been accepted provided there were astronomers supporting the corresponding theories. The arbitrary acceptance of one of them was, in fact, more reasonable than the skeptic rejection of all three arguments. In any case, the preference of $\{\alpha\}$, $\{\beta\}$ or $\{\gamma\}$ over $\emptyset$ clearly exceeds even the bounds of credulity established by sustainable semantics. That preference implies to believe something rather than nothing, even at the price of being unable to defend that belief. In the realm of science this seems to accord with Kuhn's claim that to reject a paradigm without simultaneously substituting it by another is to reject science itself (cf. [13]). Scientific skepticism prescribes to try to refute any theory but, contrary to the Popperian precept, refutations (or the existence of anomalies) do not provide reasons enough for discarding a theory if there is no better alternative to it. According to this pragmatic point of view, sustainable and, *a fortiori*, also preferred semantics seem to be excessively skeptic for representing scientific argumentation.

Cycles of attacks can also arise when several criteria are involved in a practical decision situation. Suppose, for instance, that you have to choose a school for your children. Then, you will evaluate the alternatives according to, say, three criteria: tuition fee (t), nearness (n) and social environment (e). Assume you finally reduce the options to three schools, say s_1, s_2 and s_3, which you order according to those three criteria as follows:

$$s_1 \succ_t s_2 \succ_t s_3$$

$$s_2 \succ_n s_3 \succ_n s_1$$

$$s_3 \succ_e s_1 \succ_e s_2$$

This situation poses what is known in social choice theory as the "Condorcet's paradox", i.e. the acyclicity of the global preference relation obtained aggregating the individual preferences through majority voting. More precisely, if we define the global preference relation $\succ$ as $s \succ s'$ iff $|\{c\colon\ c \in C \text{ and } s \succ_c s'\}| > |\{c\colon\ c \in C \text{ and } s \not\succ_c s'\}|$, where $C = \{t, n, e\}$, we obtain the cycle $s_1 \succ s_2 \succ s_3 \succ s_1$. This situation leads to an argumentation framework with the structure of Example 7, where the arguments are:

α: "s_2 is better than s_3 with respect to the nearness and the tuition fee, but s_1 is better than s_2 with respect to the tuition fee and the environment, so s_1 is the best choice";
β: "s_3 is better than s_1 with respect to the environment and the nearness, but s_2 is better than s_3 with respect to the nearness and the tuition fee, so s_2 is the best choice";
γ: "s_1 is better than s_2 with respect to the tuition fee and the environment, but s_3 is better than s_1 with respect to the environment and the nearness, so s_3 is the best choice".

Is in these situations of practical decisions where the choice of one argument is clearly more sensible than the rejection of them all. As far as we know, Baroni et al. have proposed the only semantics for dealing with odd-length cycles of attack

that resembles our analysis [3]. Guided by similar intuitions, our next semantics is based on accepting a set of arguments A whenever every set of arguments that challenges A ends up challenged by the last set in a sequence of challenges started by A. More precisely:

Definition 9. A strategy $A \subseteq AR$ is *cyclically cogent* iff for every strategy C, if $C \hookrightarrow A$ but $A \not\hookrightarrow C$ then there exists a sequence $D_1, \ldots, D_m$ of strategies such that $D_1 = C$, $D_m = A$ and $D_{k+1} \hookrightarrow D_k$ but $D_k \not\hookrightarrow D_{k+1}$, for $1 \leqslant k < m$.

Definition 10. $A \subseteq AR$ is a *tolerant extension* iff A is a maximal (w.r.t. $\subseteq$) cyclically cogent strategy.

Example 11 *(Example 7 revisited).* The tolerant extensions are $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$.

From Definitions 4 and 9 it follows immediately that cogent strategies are also cyclically cogent. This indicates that tolerant extensions incorporate a higher degree of credulity than sustainable extensions.

Proposition 12. *If A is a sustainable extension then there exists some tolerant extension B such that $A \subseteq B$.*

This proposition jointly with Proposition 8 immediately entails the following theorem:

Proposition 13. *Every preferred extension is contained in a sustainable extension, and every sustainable extension is contained in some tolerant extension.*

4. Sufficient conditions for agreement among semantics

A question that arises naturally is whether there exist sufficient conditions for the coincidence between these and other credulous semantics. One answer is that in *limited controversial* frameworks—those which exclude odd-length cycles of attack– both tolerant and sustainable extensions collapse into a preferred one. To make this claim precise, we first recall the definition of limited controversial frameworks:

Definition 14. (See Dung [10].) An argumentation framework is *limited controversial* iff there exists no infinite sequence of arguments $\alpha_1, \ldots, \alpha_n, \ldots$ such that $\alpha_i + 1$ is controversial with respect to α_i. An argument α is *controversial* with respect to an argument β iff α indirectly attacks and indirectly defends β. An argument β *indirectly attacks* α iff there exists a finite sequence $\alpha_0, \ldots, \alpha_{2n+1}$ such that (1) $\alpha = \alpha_0$ and $\beta = \alpha_{2n+1}$ and (2) for each i, $0 \leqslant i \leqslant 2n$, α_{i+1} attacks α_i. An argument β *indirectly defends* α iff there exists a finite sequence $\alpha_0, \ldots, \alpha_{2n}$ such that (1) $\alpha = \alpha_0$ and $\beta = \alpha_{2n}$ and (2) for each i, $0 \leqslant i \leqslant 2n$, α_{i+1} attacks α_i.

Then we have that:

Lemma 15. *For every limited controversial argumentation framework AF, every sustainable extension of AF is also a preferred extension of AF, and vice versa.*

Lemma 16. *In a limited controversial argumentation framework AF, every tolerant extension of AF is also a preferred extension of AF, and vice versa.*

We can put these two results together to state a sufficient condition for the equivalence among sustainable, tolerant and preferred extensions.

Proposition 17. *In every limited controversial argumentation framework, any subset of arguments is a sustainable extension iff it is a tolerant extension iff it is a preferred extension.*

Moreover, we can incorporate our semantics into the known, more general picture of relations among Dung's semantics. First, our Proposition 17 together with Theorem 25 in [10], which indicates that every preferred extension is a complete extension, entails that in every limited controversial argumentation framework, sustainable and tolerant extensions are also complete extensions. Second, Theorem 30 in [10] claims that all well-founded argumentation frameworks, i.e. those in which no infinite sequence of attacks occur, have a unique complete extension which is grounded, preferred and stable. Now, from Proposition 17 and the obvious fact that every well-founded argumentation framework is also limited controversial, we can conclude that the unique complete, grounded, preferred and stable extension in such frameworks is also sustainable and tolerant.

Summarizing, these results ensure that in "normal" cases the behavior of sustainable and tolerant semantics will be tantamount to that of preferred semantics, while different behaviors are restricted to the "weird" cases of odd-length cycles of attack.

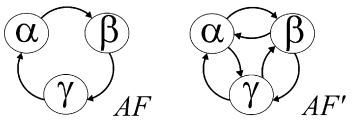

Fig. 3. Argumentation frameworks AF and AF' are equivalent under the *attack* = *conflict* + *preference* hypothesis.

5. Tolerant semantics and the hypothesis "*attacks* = *conflict* + *preference*"

In Section 3.3 we have analyzed some motivating examples of tolerant semantics from a merely intuitive point of view. Now we want to assess it through a comparison with the intuition behind preferred semantics. Our analysis will be based on the possible formal properties of the abstract concept of 'attack'. We will argue that if the intended meaning of attack involves a reflexive and transitive preference relation among arguments, the definition of tolerant semantics capture the same behaviors that arise in preferred semantics.

In his approach, Dung does not ascribe any particular formal property to the attack relation, a modeling choice that was appropriate at that stage. But it is legitimate to ask what it might represent. One possibility is that it involves at the same time two different relations: 'α attacks β' implies, on one hand, that α *disagrees* with β while, on the other hand, α is *preferred* to β. Similar points of view have been previously taken elsewhere. For example, Simari and Loui in [22] define their attack relation, called 'defeat', on the basis of a combination of *disagreement* (a symmetric relation between two arguments that together imply a contradiction in conjunction with the knowledge base) and *specificity* (which they take from [20]), which, although not a preference (it is not transitive), behaves as such in most cases of interest.

More recently, Kaci et al. [11] have shown that if the attack relation is characterized by a symmetric conflict relation plus a transitive preference relation, then a "strictly acyclic" argumentation framework obtains, where 'strictly acyclic' means (strangely) that for any cycle 'α_1 attacks α_2 attacks … attacks α_n attacks α_1' it must be the case that 'α_1 attacks α_n attacks … attacks α_2 attacks α_1'. For instance, the argumentation framework of Example 7, $AF = \langle\{\alpha, \beta, \gamma\}, \{(\alpha, \beta), (\beta, \gamma), (\gamma, \alpha)\}\rangle$, is equivalent to the argumentation framework $AF' = \langle\{\alpha, \beta, \gamma\}, \{(\alpha, \beta), (\beta, \alpha), (\beta, \gamma), (\gamma, \beta), (\gamma, \alpha), (\alpha, \gamma)\}\rangle$ under the hypothesis that *attack* = *conflict* + *preference* (see Fig. 3). Note that the tolerant extensions of AF are exactly the preferred extensions of AF'. If this happens in general and the aforementioned claim is true—i.e., for each cycle there exists a cycle on the same arguments but going in reverse direction—then we can arguably claim that tolerant semantics captures the intended behavior of preferred semantics.

Let us analyze Example 7 from an intuitive point of view, assuming that the preference relation among arguments is reflexive and transitive. Then, we can understand 'α is preferred to β' as 'α is at least as good as β'. Consequently, the cycle α attacks β attacks γ attacks α can be thought as "α, β and γ disagree among them and any one of them is at least as good as any other". This is our main point: under this interpretation it seems perfectly reasonable to choose any one of these arguments, as it is the case in tolerant extensions. Hence, tolerant extensions provide a reasonable credulous semantics for attack relations that incorporate a preference relation. The fact that tolerant extensions in this setting consist of the same subsets as the preferred extensions when the attacks in cycles go in both directions, is true except in presence of cycles of length one. Formally:

Proposition 18. *Let AF be a strictly acyclic argumentation framework free of self-attacking arguments. Then every tolerant extension of AF is also a preferred extension of AF, and vice versa.*

The following lemma will help to prove this claim as well as some further results:

Lemma 19. *Assume A is cyclically cogent. Then for any non-self-attacking argument β which attacks A, there exists some argument $\alpha \in A$ and a sequence γ_1 attacks … attacks γ_n such that $\alpha = \gamma_1$ and $\beta = \gamma_n$.*

Let AF^o be the closure of AF w.r.t. strict acyclicity, i.e. if $AF = \langle AR, attack\rangle$ then $AF^o = \langle AR, \{attack \cup \{(\alpha, \beta)\colon\ (\beta, \alpha) \in attack$ and there exists a sequence $\gamma_1, \ldots, \gamma_n$ such that γ_{i+1} attacks γ_i, $\gamma_1 = \beta$, $\gamma_n = \alpha$, $1 \leqslant i < n\}\rangle$. Provided that AF is free of self-attacking arguments, the preferred extensions of AF^o are the same as the tolerant extensions of AF. But if self-attacking arguments are present in AF, then tolerant extensions might justify more arguments than those sanctioned by preferred extensions in both AF and AF^o (cf. Example 6). This means that if we interpret the attack relation in terms of disagreement plus a reflexive and transitive preference relation, the tolerant semantics in AF is as reasonable as the preferred semantics in AF^o. Moreover, if self-attacking arguments introduce some kind of "anomaly"[3] into preferred semantics—Dung seems to adhere to this opinion (cf. [10, p. 351])—this "anomaly" is resolved by the tolerant semantics. The following results make this claim clear.

[3] We use this term with the approximate meaning in Philosophy of Science: an *anomaly* of a theory is an event that the theory should explain but does not.

Proposition 20. *The class of all the tolerant extensions of AF is exactly the class of all tolerant extensions of AF^o.*

Proposition 21. *Let AF be a strictly acyclic argumentation framework free of self-attacking arguments. Then every tolerant extension of AF is also a preferred extension of AF^o, and vice versa.*

Finally, we have to mention an alternative hypothesis about the attack relation that is due to Amgoud and Cayrol [2]. They propose that A attacks B iff $A \mathrel{R} B$ and it is not the case that $B \gg^{Pref} A$, where $R \subseteq AR \times AR$ represents a defeat relationship between arguments, and $\gg^{Pref}$ is a strict ordering associated with a preorder (reflexive and transitive) *Pref* on AR. Mutatis mutandi, here 'defeat' can be understood as synonymous of 'conflict'. Note that from this definition of 'attack' it follows that if A attacks B but not vice versa, then $A \gg^{Pref} B$. Under this interpretation, the attack relation in the argumentation framework of Example 7 yields the cycle $A \gg^{Pref} B \gg^{Pref} C \gg^{Pref} A$. As a consequence, in this and similar cases of odd-length cycles of attack we will obtain the same results as under the *attack = conflict + preference* hypothesis.

6. Sustainable and tolerant semantics in a strategic argumentation game

In this section we will show, in game-theoretical terms, how some strategic considerations that lead to the choice of a set of arguments may lay out a rational decision-making basis for sustainable and tolerant semantics.

In Game Theory, the process by which the players of a game sequentially delete from their sets of alternatives all those strategies that are dominated by any other strategy is known as *iterated elimination of weakly dominated strategies* (IEWDS). A strategy is weakly dominated if, whatever the choice of the other agents, there exists some other strategy that pays at least the same for the actual choice of the other agents, and more for at least one alternative choice of the other agents. Then, consider two individuals i and j playing a game in which they have to choose simultaneously actions. To decide which decisions to make they should apply the IEWDS process. So i first eliminates all his dominated strategies and then, believing that j is rational (and therefore that she will eliminate all her dominated strategies), will restrict his remaining strategies by ignoring the dominated actions of j. On this reduced set, i will run a round of elimination of dominated strategies, and by assuming that j will do the same, i further restricts his options, choosing only the undominated actions, and so on. In turn, without communicating with i, j will do the same. The surviving strategies in the iterated elimination process are considered "rationalizable" (see [15]).[4] We will show next a correspondence between the tolerant extensions of a framework AF and the surviving strategies of the IEWDS process in a game associated to AF.

A *strategic argumentation game* associated to an argumentation framework $AF = \langle AR, attacks\rangle$, denoted by '$SAG_{AF}$', is a game between two agents, 1 and 2. The strategies that each agent can choose are elements of 2^{AR}. The agents obtain payoffs from the outcome of the game, according to the following function:

$$Pay_i(A,B) = \begin{cases} 0, & \text{if } A \text{ is not conflict-free, or there exists an argument} \\ & \sigma \in B \text{ such that } \sigma \textit{ attacks } A \text{ but not } A \textit{ attacks } \sigma; \\ 1-x, & \text{otherwise,} \quad \text{where } x = \begin{cases} \epsilon, & \text{if } A \subset B \text{ and } Pay_{-i}(A,B) > 0; \\ 0, & \text{otherwise} \end{cases} \end{cases}$$

where $Pay_i(A,B)$ is the payoff obtained by player i when she plays strategy A and agent $-i$ plays strategy B (respectively, $Pay_{-i}(A,B)$ is the payoff obtained by $-i$ in the same situation). Note that this payoff function was defined according to the same idea of defensibility that is behind the notions of admissibility and cogency. Agent i gets 0 if either her strategy yields inconsistencies (i.e. it contains arguments that attack others within that strategy), or the chosen strategy is defenseless against a possible attack from the strategy chosen by the other agent, $-i$. Otherwise, when the strategy is conflict-free and can be defended against any attack from $-i$'s strategy, i gets a payoff of 1, unless $-i$ has chosen the same and more arguments in a defensible strategy, in which case i obtains $1-\epsilon$. The (small) constant ϵ intends to model a slight penalty for not choosing a larger class of arguments. The game is better understood as a confrontation of "theories" rather than as a debate in which one agent proposes some thesis that the other one tries to refute.

Let us consider the strategic argumentation game corresponding to the framework of Example 7 (Fig. 4). In this setting, any of the strategies $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$ is a rationalizable choice. For example, $\{\alpha\}$ is not dominated by any other strategy, because it would pay more than any other one if the rival were to choose $\{\beta\}$. This means that $\{\alpha\}$ will survive the elimination process, and hence it is rationalizable. The same can be said about $\{\beta\}$ and $\{\gamma\}$. This suggests that, in terms of argumentation frameworks, $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$ are "rationalizable" extensions. These extensions are captured by tolerant semantics.

The relation between tolerant extensions and the IEWDS process can be more technically detailed through the notion of "mixed strategy Nash equilibrium".[5]

[4] Actually, rationalizable strategies are those that survive a process of iterated elimination of *strongly* dominated strategies, i.e. those which pay strictly less than the ones that survive. IEWDS survivors are sometimes called "admissible" strategies in the Game Theory literature, but to avoid overloading the term we will call them *rationalizable* as well.

[5] The technical details that follow can be ignored by readers who are not familiar with Game Theory, although the presentation is self-contained.

		2							
		$\emptyset$	$\{\alpha\}$	$\{\beta\}$	$\{\gamma\}$	$\{\alpha,\beta\}$	$\{\beta,\gamma\}$	$\{\alpha,\gamma\}$	AR
1	$\emptyset$	$1,1$	$1-\epsilon,1$	$1-\epsilon,1$	$1-\epsilon,1$	$1,0$	$1,0$	$1,0$	$1,0$
	$\{\alpha\}$	$1,1-\epsilon$	$1,1$	$1,0$	$0,1$	$1,0$	$0,0$	$0,0$	$0,0$
	$\{\beta\}$	$1,1-\epsilon$	$0,1$	$1,1$	$1,0$	$0,0$	$1,0$	$0,0$	$0,0$
	$\{\gamma\}$	$1,1-\epsilon$	$1,0$	$0,1$	$1,1$	$0,0$	$0,0$	$1,0$	$0,0$
	$\{\alpha,\beta\}$	$0,1$	$0,1$	$0,0$	$0,0$	$0,0$	$0,0$	$0,0$	$0,0$
	$\{\beta,\gamma\}$	$0,1$	$0,0$	$0,1$	$0,0$	$0,0$	$0,0$	$0,0$	$0,0$
	$\{\alpha,\gamma\}$	$0,1$	$0,0$	$0,0$	$0,1$	$0,0$	$0,0$	$0,0$	$0,0$
	AR	$0,1$	$0,0$	$0,0$	$0,0$	$0,0$	$0,0$	$0,0$	$0,0$

Fig. 4. Three-length cycle of attack.

		2			
		$\emptyset$	$\{\alpha\}$	$\{\beta\}$	AR
1	$\emptyset$	$1,1$	$1,0$	$1-\epsilon,1$	$1,0$
	$\{\alpha\}$	$0,1$	$0,0$	$0,0$	$0,0$
	$\{\beta\}$	$1,1-\epsilon$	$0,0$	$1,1$	$0,0$
	AR	$0,1$	$0,0$	$0,0$	$0,0$

Fig. 5. The "lottery paradox" paradox strategic argumentation game.

Definition 22. Given a space of actions $S_1 \times S_2$, where S_1 and S_2 are the sets of strategies available to agents 1 and 2, respectively, a *mixed strategy Nash equilibrium* is a pair $\langle \sigma_1, \sigma_2 \rangle$ where σ_i is a probability distributions over S_i, for $i = 1, 2$ (i.e. if $S_i = \{A_1^i, \ldots, A_k^i\}$, $\sigma_i = (p_1^i, \ldots, p_k^i)$, where p_j^i is the probability of action A_j^i and $\sum_{j=1}^{k} p_j^i = 1$). Furthermore, for the agent i and every alternative distribution τ_i, $E_i(\sigma_i, \sigma_{-i}) \geqslant E_i(\tau_i, \sigma_{-i})$, where $E_i(\cdot)$ denotes the expected payoff function of agent i given her own as well as $-i$'s probability functions.

Under this characterization it is easy to see in Example 7 that if σ is a probability distribution over $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$, both $\langle \emptyset, \sigma \rangle$ and $\langle \sigma, \emptyset \rangle$ are mixed strategy Nash equilibria. To see this, consider that player i chooses $\emptyset$. Then, player $-i$ obtains the same payoff (1) on each of the pure strategies $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$. By the Fundamental Theorem of Mixed Strategy Nash Equilibria (cf. [15]), these three strategies belong to the support of a mixed strategy σ, i.e. they have non-zero probability assigned by σ. Since the expected payoff of σ is 1, σ is a best response to $\emptyset$.

Conversely, if player $-i$ chooses σ, player i compares the expected payoffs of playing any of her pure strategies. Since $\emptyset$ yields a payoff $1 - \epsilon$ while any of $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$ yields a expected payoff of $2/3$, it follows that player i yields her best response by playing $\emptyset$.[6]

Notice that this result is independent of the exact probability distribution over $\{\alpha\}$, $\{\beta\}$ and $\{\gamma\}$. The only condition on σ is that each of those arguments has to have a strictly positive probability. In the previous case, either $\{\alpha\}$, $\{\beta\}$ or $\{\gamma\}$ can be accepted in the case that they can be played out in a mixed Nash equilibrium. The same, of course is true for $\emptyset$ that can be seen as a degenerate mixed strategy in which $\emptyset$ itself has probability 1.

A strategy that becomes eliminated in an IEWDS process can never carry positive probability in a mixed strategy Nash equilibrium. This follows from the Fundamental Theorem of Mixed Strategy Nash Equilibria (cf. [15]), and implies that:

Proposition 23. *If A is a tolerant extension of AF then A survives a IEWDS process in SAG_{AF}, i.e., A has a positive probability in a mixed strategy Nash equilibrium.*

The converse is not true, which is proved by $\emptyset$ in our example: it has positive probability in a mixed strategy Nash equilibrium but it is not a tolerant extension. On the other hand, it is easy to see that if A has a positive probability in a mixed strategy Nash equilibrium, then A is contained in some tolerant extension, which is true in particular for the intersection of all the tolerant extensions.

This kind of choice also gives the expected outcome in the *lottery paradox* paradox game. The solution, obtained by running as they run a IEWDS process, is the equilibrium profile $(\{\beta\}, \{\beta\})$. This amounts to say that β has probability 1, while each other strategy is assigned probability 0. In contrast with the case of the three-cycle of attacks, $\emptyset$ is not a good choice here since it is eliminated by $\{\beta\}$ after the elimination of $\{\alpha\}$. This outcome is exactly the same as prescribed by the sustainable semantics (see Fig. 5).

In general, an agent that plays a sustainable extension proceeds by eliminating all the strategies A such that for some strategy B, $Pay_i(A, B) = 0 < Pay_i(B, A)$, and then arbitrarily selecting any non-weakly dominated strategy among the survivors. In Example 7, $\emptyset$ is the only strategy that survives the elimination and is non-weakly dominated.

[6] If $\sigma = (p_{\{\alpha\}}, p_{\{\beta\}}, p_{\{\gamma\}})$ is the strategy used by i, the expected payoffs that $-i$ gets by playing either one of her pure strategies $\{\alpha\}$, $\{\beta\}$ or $\{\gamma\}$ are $p_{\{\alpha\}} + p_{\{\beta\}}$, $p_{\{\beta\}} + p_{\{\gamma\}}$ and $p_{\{\alpha\}} + p_{\{\gamma\}}$, respectively (for example, if $-i$ plays $\{\alpha\}$ then her expected payoff is $1p_{\{\alpha\}} + 1p_{\{\beta\}} + 0p_{\{\gamma\}} = p_{\{\alpha\}} + p_{\{\beta\}}$). Since each of $p_{\{\alpha\}}$, $p_{\{\beta\}}$ and $p_{\{\gamma\}}$ are strictly positive and $p_{\{\alpha\}} + p_{\{\beta\}} + p_{\{\gamma\}} = 1$, the payoff for $-i$ of playing either one of $\{\alpha\}, \{\beta\}$ or $\{\gamma\}$ is strictly less than 1.

More formally, A is cogent if only if $A \in Cog = \{A\colon$ for every B, $B \subseteq AR$, if $Pay_i(B, A) \geqslant 1 - \epsilon$ then $Pay_i(A, B) \geqslant 1 - \epsilon\}$. Then:

Lemma 24. *For every $A, B \in Cog$, $Pay_i(A, B) \geqslant 1 - \epsilon$.*

Lemma 25. *If Cog is finite then there exists some $A \in Cog$ such that for every $B \in Cog$, $Pay_i(A, B) = 1$.*

Finally,

Proposition 26. *If A is a sustainable extension of AF then A is not weakly dominated in SAG_{AF} restricted to the strategies in Cog (the converse is true just if Cog is finite).*

Contrary to the choice of a tolerant extension, the choice of sustainable extensions yields not mixed but pure strategies equilibria. Note that if both players choose sustainable extensions they avoid the risk of obtaining less than 1. In this sense, higher degrees of credulity in the extension semantics of an argumentation framework correspond to facing higher risks of loss in the argumentation game.

7. Discussion and related work

7.1. Comparison with Baroni et al.'s CF2 semantics

Let us discuss now some alternative approaches. Notions similar to tolerant semantics have been introduced by Baroni and Giacomin in [3] and Baroni et al. in [4]. Nevertheless, their semantics are not totally equivalent to ours, although we find difficult to say which one is more intuitive. Anyway, let us see the formal description of their semantics, called *CF2*, in order to show the difference. The approach is inspired in the graph-representation of the relation of attack among arguments. In this view, an argumentation framework AF can be decomposed along its strongly connected components, $SCCS_{AF}$. Every strongly connected component $SCC \in SCCS_{AF}$ is a set of arguments, all of which are path-equivalent. This relation is reflexive, and two different arguments α and β are path-equivalent iff there is a path from α to β and there is a path from β to α. Since the SCC's can be ordered following the direction of the graph, the original idea is that the directionality can be used to yield recursively the semantics. They define a generic recursive function, for which the base case is ad-hoc for each semantics. The generic function for the semantics we are considering is defined as follows: $E \in \mathcal{GF}(AF, C)$:

- in case $|SCCS_{AF}| = 1$, $E \in \mathcal{BF}(AF, C)$,
- otherwise $\forall S \in SCCS_{AF}(E \cap S) \in \mathcal{GF}(AF \downarrow_{UP(S,E)}, U(S, E) \cap C)$,

where $\mathcal{BF}(AF, C)$ is a base function such that given $|SCCS_{AF}| = 1$ and a set $C \subseteq AR$, yields a maximal conflict-free subset of AR. C is the set of nodes which have no attackers from outside S or whose outer attackers are all attacked by E. $U(S, E)$ is the subset of nodes of S that are not attacked by E from outside S and are defended by E (i.e., their attackers from outside S are all attacked by E). And $UP(S, E)$ consists of $U(S, E)$ joint with the set of all the provisionally defeated nodes, that is, the nodes of S that are not attacked by E from outside S and are not defended by E (i.e., at least one of their attackers from outside S is not attacked by E). This semantics sanctions the same extensions as tolerant semantics in the frameworks of Examples 6 and 7. But in the following example the behavior is different:

Example 27. Let $AF = \langle\{\alpha, \beta, \gamma, \delta, \epsilon, \eta, \phi\}, \{(\alpha, \beta), (\beta, \gamma), (\gamma, \alpha), (\alpha, \delta), (\delta, \epsilon), (\epsilon, \eta), (\eta, \phi), (\gamma, \phi), (\phi, \delta)\}\rangle$. Then $\{\alpha, \eta\}$ is a tolerant extension but it is not an extension in the *CF2* semantics. In order to see that it is a tolerant extension note that for any subset of arguments B such that $B \hookrightarrow \{\alpha, \eta\}$, there exists a sequence of subsets of arguments $D_1 \hookrightarrow \cdots \hookrightarrow D_m$, where $D_1 = \{\alpha, \eta\}$ and $D_m = B$. Such subsets are $\{\alpha, \epsilon\}$, $\{\gamma, \epsilon\}$, $\{\gamma, \delta, \eta\}$, $\{\gamma, \eta\}$ and $\{\alpha, \epsilon, \phi\}$, for which we found the following sequences:

(1) $\{\alpha, \eta\} \hookrightarrow \{\beta, \delta, \eta\} \hookrightarrow \{\beta, \phi, \epsilon\} \hookrightarrow \{\gamma, \delta, \eta\} \hookrightarrow \{\gamma, \epsilon\} \hookrightarrow \{\alpha, \epsilon\}$;
(2) $\{\alpha, \eta\} \hookrightarrow \{\beta, \delta, \eta\} \hookrightarrow \{\gamma, \eta\}$;
(3) $\{\alpha, \eta\} \hookrightarrow \{\beta, \delta, \eta\} \hookrightarrow \{\beta, \phi, \epsilon\} \hookrightarrow \{\alpha, \epsilon, \phi\}$.

Sequence 1 completes the cycles for $\{\alpha, \epsilon\}$, $\{\gamma, \epsilon\}$ and $\{\gamma, \delta, \eta\}$, sequence 2 completes the cycle for $\{\gamma, \eta\}$ and sequence 3 completes the cycle for $\{\alpha, \epsilon, \phi\}$. On the other hand, to see that $\{\alpha, \eta\}$ is not sanctioned by the *CF2* semantics, note that we have two SCC's: $S_1 = \{\alpha, \beta, \gamma\}$ and $S_2 = \{\delta, \epsilon, \eta, \phi\}$. Assume $E \cap S_1 = \{\alpha\}$, then for S_2 the parameters of the generic function are $AF \downarrow_{\{\epsilon,\eta,\phi\}}$ and $U(S_2, E) \cap C = \{\epsilon, \phi\}$. Hence, $SCCS_{AF\downarrow_{\{\epsilon,\eta,\phi\}}} = \{\{\epsilon\}, \{\eta\}, \{\phi\}\}$. Taking the strongly connected component $S' = \{\eta\}$, we have $(E \cap S') \in \mathcal{GF}(AF \downarrow_{\emptyset}, \emptyset) = \emptyset$, since $UP(S', E) = \emptyset$ and $C = \emptyset$. Therefore, $\{\alpha, \eta\}$ cannot be an extension in this semantics.

Despite this difference, both tolerant and *CF*2 semantics display similar behaviors in many cases.

Sustainable semantics, on the other hand, is an alternative for dealing properly with the problem of self-attacking arguments, overruling arguments belonging to larger odd-length cycles of attack. We have not found any other approach yielding similar results. As we have seen, this semantics offers a solution to both problems, agreeing with Pollock on how to deal with the interpretations of Example 7 [17,18]. On the other hand, the solution given by Pollock to the problem of self-attacking arguments in [16] and [17] consists in inhibiting *by definition* their interference in the warrant process. To be fair, Pollock changed his view later [19] by accepting that self-defeat is really a necessary feature for his non-monotonic entailment system. Our solution, instead, inhibits the interference of self-attacking arguments as an emergent ability of the model, not by fiat. Of course, it contributes the fact that we define a credulous semantics instead of a skeptical one.

7.2. Comparison with Jakobovits and Vermeir's robust semantics

Despite the similar treatment given to self-attacking arguments by tolerant, sustainable and *CF*2 semantics, the question cannot be declared settled. The lack of indisputably sound examples contributes to maintain the fuzziness of the concepts and the disagreement among authors. In [14], for instance, Jakobovits and Vermeir provide an extended version of Example 6, where a third argument, say γ, is attacked by β. In their *robust semantics*, β must be disabled by the attack from α, while γ must be reinstated. This is clearly at odds with the generally accepted interpretation—to which we adhere—according to which α must be overruled, β reinstated and γ defeated by β. On the other hand, their semantics fits well to the example presented by them, which takes the arguments from the logic program:

$$p \leftarrow \neg p$$

$$q \leftarrow \neg p$$

mapping proof trees into arguments in the following way:

$$\alpha : -\neg p - p$$

$$\beta : -\neg p - q$$

$$\gamma : -\neg q$$

where α attacks α attacks β attacks γ (an attack occurs when the conclusion of a tree contradicts some leaf node of the other tree). By a labeling process defined by them, '$\pm$' is assigned to α (α's conclusion takes an undefined value), '$-$' to β (the negation of β's conclusion is accepted) and '$+$' to γ (accepts γ's conclusion). Since q cannot be founded using $\neg p$, then $\neg q$ may be accepted using negation as failure. The proposed semantics seems right in the light of this example. Nevertheless, consider the following example:

$$q \leftarrow \neg p$$

$$p \leftarrow q$$

where by mapping proof trees into arguments we obtain:

$$\alpha : -\neg p - q - p$$

$$\beta : -\neg p - q$$

Although this has the structure of the *lottery paradox* paradox, β, which is a sub-argument of α, is assigned '$-$' and, hence, rejected, against our intuition. An open question that deserves a deep philosophical analysis is to determine whether a self-attack by a sub-argument is a special case to deal with or it should be treated as any other self-attack.

7.3. Strategic argumentation games and dialogue games

The strategic view that justifies our notion of cogency is not the only possible "two-agents" point of view on argumentation. Other approaches focused instead on *dialogue* games. Among the salient works along this line we can mention those of Vreeswijk [23], Prakken [21] and Vreeswijk and Prakken [24]. The latter is particularly close to the present work since it deals with the problem of representing credulous and skeptic argumentation, along the intuitions of Dung. These authors postulate a proof-theoretic procedure based on a specific dialogue game they call TPI-*dispute* (for *two-party immediate response dispute*). This game assumes two players, PRO and CON, and rules governing the players' choices:

(1) PRO can repeat itself if possible, because CON might fail to find a new attacker against PRO's repeated argument.
(2) CON should repeat PRO (if possible), because it would prove that PRO's collection of arguments is not conflict-free.
(3) PRO should not repeat CON, because that would introduce a conflict in PRO's own collection of arguments.
(4) CON should not repeat itself unless the second use is in a different line of the dispute.

A TPI-dispute is carried out through two kinds of moves:

- *block*: to prompt an argument that places the other player in a position in which she cannot move.
- *eo ipso*: to prompt a previous non-backtracked argument of the other player, in order to show a conflict within her collection of arguments.

The main results in [24] are the following. An argument can be defended by PRO in every TPI-dispute if and only if it is in some preferred extension. This stands for credulous argumentation. On the other hand, in systems where all preferred extensions are also stable, an argument is in every preferred extension if and only if it can be defended by PRO in every TPI-dispute and none of its attackers can be so. This, in turn, stands for skeptic argumentation. Besides the obvious differences between Vreeswijk and Prakken's approach and ours, we find more subtle distinctions in the strategic analysis assumed in each approach. From a dialogical point of view, strategies can be seen as indicating which (single) argument has to be uttered at each circumstance. Furthermore, to know whether an argument is either credulously or skeptically justified requires to play each possible game. Game theory, instead, allows to represent *all* the potential confrontations in just a single matrix (no matter how large), and consequently the characterization of the epistemic attitudes obtains through suitable solution concepts.

Tolerant semantics can be captured by introducing the following "rule of tolerance": accept all the arguments α such that for any β, if β is a blocking argument in a TPI-dispute in which PRO advances α, then β cannot be defended in a TPI-dispute. This amounts to say that the burden of proof is on the attacker, not on the defender. Taking into account Lemma 19 the following result can be obtained:

Proposition 28. *An argument α belongs to a tolerant extension iff no argument β that blocks α in a* TPI*-dispute can be defended in a* TPI*-dispute.*

In the *lottery paradox* paradox (Example 6), β is accepted according to the "rule of tolerance" because α, its blocker, cannot be defended in a TPI-dispute (even when β cannot be defended neither):

Dispute 1.

1. PRO: β
2. CON: α (block, CON wins)

(CON blocks PRO with α because PRO cannot repeat CON's move advancing α in time).

Dispute 2.

1. PRO: α
2. CON: α! (*eo ipso*, CON wins)

(α cannot be accepted according to the "rule of tolerance" because α, its (self) attacker, is not introduced by CON in a blocking move, but in an *eo ipso* move).

In the three-length cycle of attacks of Example 7, all the three arguments can be justified because there exist no winning strategies for their respective attackers. The following dispute shows that β cannot be defended, hence γ (which is attacked only by β) must be accepted according to the "rule of tolerance".

Dispute 3.

1. PRO: β
2. CON: α
3. PRO: γ
4. CON: γ! (*eo ipso*, CON wins).

Sustainable strategies can be captured in a TPI-like dialogue by changing rule 3 to allow PRO to repeat CON's arguments, only as a way of pointing out a conflict within CON's strategy (*eo ipso*). Let us call such disputes 'TPI*-disputes'. They are complete and sound with respect to sustainable semantics:

Proposition 29. *An argument is in a sustainable extension iff it can be defended in every* TPI**-dispute.*

As an example, we present two TPI*-disputes on the *lottery paradox* paradox: Dispute 4 shows a successful defense of argument β (the only sustainable solution) while Dispute 5 shows a failed defense of argument α:

Dispute 4. [TPI*]

1. PRO: β
2. CON: α
3. PRO: α! (*eo ipso*, PRO wins)

(PRO is allowed to repeat CON's move just for an *eo ipso* move).

Dispute 5. [TPI*]

1. PRO: α
2. CON: α! (*eo ipso*, CON wins)

(equivalent to a TPI-dispute).

7.4. Odd-length cycles of attack in Bench-Capon's value-based argumentation frameworks

In [6], Bench-Capon introduces value-based argumentation frameworks, which give a turn to the problem of odd-length cycles of attack. On the basis of Dung's argumentation frameworks, this author considers a non-empty set of values V, and a function *val* which maps elements of *AR* on elements of V. An argument A gets a value v, $v \in V$, if accepting A promotes or defends value v. Another element considered in the model is the *audience* to which the arguments are addressed. An audience a is identified with a strict preference relation $Valpref_a$ (irreflexive, asymmetric and transitive) among the values in V. Then, besides the attack relation among arguments, Bench-Capon constructs a relation of *defeat for an audience*: an argument A *defeats an argument* B *for audience* a if and only if $attacks(A, B)$ and $val(B)$ is not more preferred (in $Valpref_a$) than $val(A)$. The concept of *conflict-free* sets of arguments is then made relative to the audience. Given $A, B \in S$, an attack from A on B does not make S conflictive for an audience if, in addition, A does not defeat B for that audience. The notions of *acceptable*, *admissible* and *preferred extension for an audience* a obtain by substituting the attack relation in the corresponding Dung's definitions by Bench-Capon's defeat.

Let us consider again the framework of Example 7, assuming now that we have two values $V = \{red, blue\}$, a valuation $val(\alpha) = red$, $val(\beta) = val(\gamma) = blue$ and two audiences r and b identified with $Valpref_r(red, blue)$ and $Valpref_b(blue, red)$, respectively (example taken from [6]). Then for audience r, which prefers the value *red* to the value *blue*, $\{\alpha, \gamma\}$ is a preferred extension, while $\{\alpha, \beta\}$ is a preferred extension for audience b, which prefers *blue* to *red*. Moreover, α is "objectively acceptable" in the example because it is preferred for every audience. Clearly, every tolerant extension in that framework is a subset of some preferred-for-an-audience extension, which lead us to interpret tolerant extensions as sets of "subjectively acceptable" arguments for some possible valuations and audiences. Bench-Capon's value-based argumentation frameworks, like our own approach, challenges the usual intuition that rejects the acceptance of arguments involved in odd-length cycles of attack. The difference is that in Bench-Capon's system the problem of acceptance is derived from considerations on the defeat-for-an-audience relation. In this way, the problems raised from odd-cycles of attack disappear. On the other hand, cycles in the defeat-for-an-audience relations are avoided by definition since they are based on preferences conceived as strict orders.

7.5. Controversial arguments in prudent semantics

Another issue to discuss is the occurrence of indirect attacks inside extensions. Sustainable and tolerant semantics cannot avoid such occurrences. Consider the following example taken from Coste-Marquis et al. [9]: $AF = \langle\{\alpha, \beta, \gamma, \delta, \epsilon, \nu\}, \{(\gamma, \beta), (\beta, \alpha), (\gamma, \epsilon), (\epsilon, \delta), (\delta, \alpha), (\nu, \delta)\}\rangle$. Then $\{\gamma, \nu, \alpha\}$ is the only extension for any of Dung's semantics. It is also the only sustainable and tolerant extension. Note that γ is controversial w.r.t. α because of the sequence γ attacks ϵ attacks δ attacks α (recall that an argument α is controversial w.r.t. β iff α indirectly attacks and defends β). Coste-Marquis et al. define a set of arguments S as *p(rudent)-admissible* iff every argument that belongs to S is acceptable in S and there are no odd-length paths of attacks connecting two arguments that belong to S (i.e., S must be admissible and free of controversial arguments), and a *p-preferred extension* is a maximally p-admissible set of arguments. So the only prudent preferred extension in the example above is $\{\gamma, \nu\}$. As this example shows, prudent semantics accepts less arguments than sustainable and tolerant semantics.

A natural way of obtaining "prudent" sustainable and tolerant extensions is to introduce the notions of *p-cogent* and *p-cyclically cogent* sets of arguments, substituting admissibility by p-admissibility in the definitions of 'cogent' and 'cyclically cogent', respectively. In this way, no controversial argument would belong to p-sustainable and p-tolerant extensions.

Cayrol et al. [8] have recently extended prudent semantics to bipolar argumentation frameworks. Besides the attack relation among arguments, the authors introduce another kind of interaction that they call 'support'. A supporting argument

enters a debate when "an agent brings to light some new piece of information and so advances an argument which justifies an assumption used by an argument provided by another agent (agents are assumed independent). This kind of interaction between arguments is not captured by the notion of defense. It is rather a kind of support" (p. 262). So the authors define bipolar argumentation frameworks $BAF = \langle AR, R_{att}, R_{sup}\rangle$ where R_{att} is the usual, Dung's style attack relation, and R_{sup} is the relation of support among arguments. To cope with the problem of controversial arguments, the authors indicate that S is a *b*(*ipolar*)*p*(*rudent*)*-conflict-free* set of arguments iff $\nexists \alpha, \beta \in S$ s.t. there exists a sequence $\alpha_1 R_{sup} \ldots R_{sup} \alpha_n R_{att} \ldots R_{att} \alpha_{n+m}$, $n = 1$, with $\alpha_1 = \alpha$, $\alpha_{n+m} = \beta$, and m is an odd number. That is, any argument α that supports (directly or indirectly) an argument α' that attacks (directly or indirectly) some argument β, is considered to be in conflict with β. *bp*-admissible sets are those *bp*-conflict-free sets such that all the arguments belonging to them are acceptable. Then the *bp*-versions of preferred and stable semantics obtain in the obvious way. But the results are not so obvious, since the new semantics elicits some surprising behaviors. For instance, some non-attacked arguments, which always belong to every preferred extension, may not belong to some *bp*-preferred extension. In comparison with tolerant semantics, it is clear that no argument belonging to odd-length cycles of attack will be warranted by a *bp*-semantics since such arguments are always controversial. On the other hand, in the framework $BAF = \langle AR, \{(\alpha, \alpha), \{(\alpha, \beta)\}, \emptyset\rangle$, argument β, which is warranted by sustainable semantics, is not warranted by any *bp*-semantics.

7.6. Further work

More work has to be done to relate sustainable and tolerant semantics with other approaches. Besnard and Doutre [5] have introduced an equational characterization of Dung's stable and complete extensions, although the authors have not found appropriate, effective equations characterizing preferred and grounded extensions. The same work could be carried out for sustainable and tolerant extensions.

Another point for future exploration is the behavior of rule-based systems (e.g. [22]) in the light of our semantics, assessing them in the light of the rationality postulates of consistency and closure recently introduced by Caminada and Amgoud [7].

8. Conclusion

In the final section of his paper, Dung recognizes that his approach lacks a solution for the problem of self-attacking arguments [10, p. 351]. In this paper we intended to give two alternative semantics, that we deemed sustainable and tolerant extensions, to solve the problems of odd-length cycles of attack while keeping a close contact with Dung's work. We did this using the same basic tool of Dung: the concept of admissibility. The novelty of our approach raises from a strategy-based concept of 'cogency', which basically is a defensibility notion build upon admissibility, albeit slightly weaker than that. This conception builds upon the following ideas:

- a subset of arguments is a possible argumentation strategy of an agent,
- the admissibility of a strategy is relative to the possible strategies of the rival,
- the choice consists of the maximal strategies that can be successfully defended against those challenging strategies which in turn can be successfully defended against them (i.e. choosing those strategies that are capable of "shifting the burden of proof").

The resulting semantics were shown to be more credulous (i.e. justifying more arguments) than preferred semantics, while sufficient conditions were established for the equivalence with this and other Dung's semantics. Moreover, a game-theoretic foundation was provided showing the strategical rationale of choosing sustainable and tolerant sets of arguments in a debate.

Acknowledgements

We thank two anonymous referees for important suggestions and remarks which notoriously improved the paper. This work was partially supported by ANPCyT (Argentine National Agency of Promotion of Scientific and Technological Research) and Universidad Nacional del Sur, Argentina, as part of the projects PICTO 731 and PICT 693.

Appendix A. Proofs

Proposition 8. *If A is a preferred extension then there exists some sustainable extension B such that $A \subseteq B$.*

Proof. Assume A is a preferred extension. If A is also a sustainable extension then the proposition is proved, so let us assume it is not. By definition of sustainable extension we have two possible cases: either A is not cogent (i.e. there exists some strategy B such that $B \hookrightarrow A$ but $A \not\leftrightarrow B$) or A is not maximal (i.e. there exists some sustainable extension B such that $A \subset B$). But, while the former case contradicts Proposition 3, the latter yields the desired result. □

Proposition 12. *If A is a sustainable extension then there exists some tolerant extension B such that $A \subseteq B$.*

Proof. Assume that A is a sustainable extension. Since A is cogent it is also cyclically cogent. Now, let us suppose that A is not a maximal strategy verifying this condition. But then there exists a maximal cyclically cogent strategy containing A, which implies that there exists a tolerant extension containing A. □

Lemma 15. *For every limited controversial argumentation framework AF, every sustainable extension of AF is also a preferred extension of AF, and vice versa.*

Proof. Assume AF is limited controversial.

- ($\Rightarrow$) Suppose that A is sustainable but not preferred. Since A is conflict-free, then either (i) there exists some $\alpha \notin A$ which is acceptable w.r.t. A, or (ii) A is not admissible. Assume case (i), then $A \cup \{\alpha\} \hookrightarrow A$ but $A \not\hookrightarrow A \cup \{\alpha\}$. Contradiction. Consider case (ii), then there exists some $\alpha \in A$ which is not acceptable w.r.t. A. Then there exists a β which attacks α and A does not attack β. Since AF is limited controversial, there cannot be self-attacking arguments. This implies that for some conflict-free set B, $\beta \in B$ and $B \hookrightarrow A$ but $A \not\hookrightarrow B$. This contradicts that A is cogent, therefore, it also contradicts that is sustainable.
- ($\Leftarrow$) Assume A is preferred but not sustainable. Then there exists some B such that (iii) $B \hookrightarrow A$ but (iv) $A \not\hookrightarrow B$. From Theorem 33(1) in [10] we have that (v) A is stable (see Definition 1). (Note that (iii), (iv) and (v) together imply that $B \setminus A \neq \emptyset$.) Now, (iii) and (iv) together imply that for some $\beta \in B \setminus A$, either (a) β is acceptable w.r.t. A, or (b) β attacks A but A does not attack β. But both (a) and (b) contradict the fact that A attacks β, which is ensured by (v). □

Lemma 16. *In a limited controversial argumentation framework AF, every tolerant extension of AF is also a preferred extension of AF, and vice versa.*

Proof. ($\Rightarrow$) Let AF be a limited controversial argumentation framework and A be a tolerant extension of AF. By contradiction, assume A is not a preferred extension in AF. Then we have two cases:

- There exists some argument α belonging to A which is not acceptable w.r.t. A. It follows that there exists some argument β_1 attacking α, but A does not attack β_1. So we have that $(A \setminus \{\alpha\}) \cup \{\beta_1\} \hookrightarrow A$ but not vice versa. Since A is tolerant we know that there must be a sequence in the relation $\hookrightarrow$ leading from A back to $(A \setminus \{\alpha\}) \cup \{\beta_1\}$. In particular, there exists some argument β_2 attacking β_1, such that $(A \setminus \{\alpha\}) \cup \{\beta_2\} \hookrightarrow (A \setminus \{\alpha\}) \cup \{\beta_1\}$. Therefore, there exists a sequence '… *attacks* β_n *attacks* … *attacks* β_1 *attacks* β_0' such that $\beta_n \notin A$, $\beta_0 = \alpha$.
  Now, in order to show a contradiction with the fact that AF is limited controversial, let us prove that there exists some odd-length cycle in that sequence. We will do this in two steps: the first is to find a cycle, and the second to show that its length is odd.
  - *Existence of a cycle*: Assume there are no cycles. Then the sequence is composed by an infinite number of different arguments. Consider the subset $\{\beta_{2i}\}_{i=0}^{\infty}$: it is an admissible, conflict-free set, neither of which arguments attacks nor is attacked by $A \setminus \{\alpha\}$. Note that α could not attack $\{\beta_{2i}\}_{i=0}^{\infty}$, because AF is limited controversial. But we can face two possible cases:
    - $\{\beta_{2i}\}_{i=0}^{\infty}$ attacks α. Then $A \setminus \{\alpha\} \cup \{\beta_{2i}\}_{i=0}^{\infty}$ is a tolerant extension (since no subset of arguments is at least as cogent as it but not vice versa) and it is at least as cogent as A but not vice versa. Hence A is not a tolerant extension. Contradiction.
    - $\{\beta_{2i}\}_{i=0}^{\infty}$ does not attack α. Then $A \cup \{\beta_{2i}\}_{i=0}^{\infty}$ is a tolerant extension larger than A. Contradiction.

    Therefore, there must exist a cycle in the sequence '… *attacks* β_n *attacks* … *attacks* β_1 *attacks* β_0'.
  - *Odd length of the cycle*: Let us prove now that the above cycle has an odd length. Assume on the contrary that it has even length. Let 'β_{i+2j} *attacks* … *attacks* β_i' be that cycle, with $\beta_i = \beta_{i+2j}$, for some i, j, $0 \leqslant i < j$. Then the subset of all the arguments in the cycle which indirectly defend β_i is an admissible set, neither of which arguments attacks nor is attacked by A. Hence this set together with A constitutes an admissible set larger than A (the same happens w.r.t. the subset of all the arguments in the cycle which indirectly attack β_i). But this is a tolerant extension larger than A. Contradiction. Hence the cycle in the sequence '… *attacks* β_n *attacks* … *attacks* β_0' has an odd length, which implies that every argument β_{k+1} in the cycle is controversial w.r.t. β_k. Therefore, the argumentation framework is not limited controversial, contradicting the hypothesis.
- There exists some argument α which does not belong to A but is acceptable w.r.t. A. Let $B = \{\alpha\colon\ \alpha \notin A$ and α is acceptable w.r.t $A\}$. Then $A \cup B$ is a tolerant extension which properly contains the tolerant extension A. Contradiction.

In consequence, A is a preferred extension.

($\Leftarrow$) Assume A is a preferred extension in a limited controversial argumentation framework AF. Theorem 33(1) in [10] states that every preferred extension in a limited controversial argumentation framework is also stable, that is, it attacks

every argument not in it. But since every preferred extension is a subset of some tolerant extension (as implied by Proposition 12), if B is a tolerant extension such that $A \subset B$ then A attacks $B \setminus A$. But then B is not conflict-free, which is a necessary condition of tolerant extensions. Contradiction. □

Lemma 19. *Assume A is cyclically cogent. Then for any not self-attacking argument β which attacks A, there exists some argument $\alpha \in A$ and a sequence γ_1 attacks ... attacks γ_n such that $\alpha = \gamma_1$ and $\beta = \gamma_n$.*

Proof. Assume that A is cyclically cogent and β attacks A. Let β be not self-attacking. By the absurd, assume that there is no sequence A attacks ... attacks β. Then for any sequence ... attacks γ_n attacks ... attacks γ_1 attacks γ_0 where $\gamma_0 = \beta$ ($n \geqslant 0$), there exists γ_i such that $\{\gamma_i\} \hookrightarrow \cdots \hookrightarrow A$. But since there is no sequence A attacks ... attacks γ_i, then A is not cyclically cogent. Absurd. □

Proposition 18. *Let AF be a strictly acyclic argumentation framework free of self-attacking arguments. Then every tolerant extension of AF is also a preferred extension of AF, and vice versa.*

Proof. Assume AF is strictly acyclic and free of self-attacking arguments.

($\Rightarrow$) Let A be a tolerant extension not attacked by self-attacking arguments. Suppose by the absurd that A is not preferred. Then we have two cases:

- There exists some argument α belonging to A which is not acceptable w.r.t. A, that is, there exists β such that β attacks α but A does not attack β. From the proof of Lemma 16, we know that there must exists a sequence of attacks going from α to β, closing a cycle. But since AF is strictly acyclic, then there also exists a cycle in the inverse direction, hence α attacks β. This implies that A attacks β, therefore α is acceptable in A. Contradiction.
- There exists some argument α not belonging to A which is acceptable w.r.t. A. Then $A \cup \{\alpha\}$ is a cyclically cogent strategy greater than A. This contradicts that A is tolerant.

Hence, A is a preferred extension.

($\Leftarrow$) Let A be a preferred extension. Suppose by the absurd that A is not tolerant. Since Proposition 3 ensures that $A \hookrightarrow B$ for every strategy B, the only case that deserves consideration is the existence of some B such that $A \subset B$ and B is tolerant. In this case there exists some argument α which is not acceptable w.r.t. A and $\alpha \in B \setminus A$. Then there exists some β attacking α but A does not attack β. Since α belongs to a tolerant extension, from Lemma 19 we know that there exists a sequence $\gamma_1, \ldots, \gamma_n$ such that γ_{i+1} attacks γ_i, $\gamma_1 = \beta$, $\gamma_n = \alpha$ and $1 \leqslant i < n$. This sequence close a cycle, and since AF is strictly acyclic, then we also have those attacks the way around. Hence, α attacks β, which means that α can defend itself. This implies that $A \cup \{\alpha\}$ is an admissible set containing properly a preferred extension (A). This contradicts the definition of preferred extension. Hence, A is a tolerant extension.

Proposition 20. *The class of all the tolerant extensions of AF is exactly the class of all tolerant extensions of AF^o.*

Proof. ($\subseteq$) Assume A is tolerant in AF but not in AF^o. Then we have two cases:

- $A \subset B$ and B is cyclically cogent in AF^o. Assume $\beta \in B$ and $\beta \notin A$. For every γ such that γ attacks β, either γ is self-attacking and, hence, $A \cup \{\gamma\}$ is cyclically cogent in AF which is absurd, or—by Proposition 19—there exists a sequence $B \hookrightarrow \cdots \hookrightarrow \{\gamma\}$ in AF^o. This implies that there exists $\delta \in B$ such that $\delta = \delta_1$ attacks ... attacks $\delta_n = \gamma$ in AF^o. Then we have two sub cases:
  - $\delta = \delta_1$ attacks ... attacks $\delta_n = \gamma$ in AF—because of AF^o being the closure of AF w.r.t. strict acyclicity. Hence B is cyclically cogent in AF. This contradicts that its proper subset A is tolerant in AF;
  - $\gamma = \delta_n$ attacks ... attacks $\delta_1 = \delta$ in AF, hence $A \cup \{\delta\}$ is cyclically cogent in AF. This also contradicts that A is tolerant in AF.
- A is not cyclically cogent in AF^o. Then there exists some strategy B, $B \hookrightarrow A$ but there is no sequence $A \hookrightarrow \cdots \hookrightarrow B$ in AF^o. In particular, it is not the case that $A \hookrightarrow B$, which implies that there exist $\alpha \in A$ and $\beta \in B$, such that β attacks α and A does not attack β. Then:
  - β attacks α and A does not attack β in AF, which contradicts that A is cyclically cogent in AF; or
  - α attacks β (the reason why β attacks α in AF^o), which contradicts that A does not attack β in AF^o.

($\supseteq$) Assume A is tolerant in AF^o but not in AF. Then we have two cases:

- $A \subset B$ and B is cyclically cogent in AF. Assume $\beta \in B$ and $\beta \notin A$. Then for every γ such that γ attacks β, there exists a sequence $B \hookrightarrow \cdots \hookrightarrow \{\gamma\}$ in AF. Hence there also exists some $\delta \in B$ such that $\delta = \delta_1$ attacks ... attacks $\delta_n = \gamma$ in AF. But then $\delta = \delta_1$ attacks ... attacks $\delta_n = \gamma$ occurs in AF^o too. This implies that B is cyclically cogent in AF^o contradicting that its proper subset A is tolerant in AF^o.

- A is not cyclically cogent in AF. This implies that there exists some B such that $B \hookrightarrow A$ but there is no sequence $A \hookrightarrow \cdots \hookrightarrow B$ in AF. Hence $A \not\leftrightarrow B$, which means that there exist $\alpha \in A$ and $\beta \in B$ such that β attacks α and A does not attack β. Then β attacks α in AF^o, but since A is tolerant in AF^o, there exists $\gamma \in A$ such that γ attacks β in AF^o. Since A does not attack β in AF then γ does not attack β in AF, hence there must be a cycle of attacks in AF involving the attack of β over γ. In consequence there exists a sequence γ attacks ... attacks β in AF which entails that there also exists a sequence $A \hookrightarrow \cdots \hookrightarrow B$. Contradiction. □

Proposition 21. *Let AF be a strictly acyclic argumentation framework free of self-attacking arguments. Then every tolerant extension of AF is also a preferred extension of AF^o, and vice versa.*

Proof. Immediate from Propositions 18 and 20. □

Lemma 24. *For every $A, B \in Cog$, $Pay_i(A, B) \geqslant 1 - \epsilon$.*

Proof. Assume $Pay_i(A, B) = 0$. Then since $B \in Cog$, $Pay_i(B, A) = 0$. Since both A and B are conflict-free, we have a mutual attack. That means that there exists some $\alpha \in A$ such that α attacks B but B does not attack α, and there exists some $\beta \in B$ such that β attacks A but A does not attack β. But then $\{\alpha, \beta\}$ attacks both A and B but not vice versa. This contradicts the fact that A and B are cogent. □

Lemma 25. *If Cog is finite then there exists some $A \in Cog$ such that for every $B \in Cog$, $Pay_i(A, B) = 1$.*

Proof. Assume, to the contrary, that for every $A \in Cog$ there exists some $B \in Cog$ such that $Pay_i(A, B) = 1 - \epsilon$. By definition of the payoffs we have that $A \subset B$. The same is true for B, etc. This implies that there exists a sequence $A \subset B \subset \cdots \subset A$, which is absurd. □

Proposition 26. *If A is a sustainable extension of AF then A is not weakly dominated in SAG_{AF} restricted to the strategies in Cog (the converse is true if Cog is finite).*

Proof. ($\Rightarrow$) Assume A is a sustainable extension of AF, and suppose that there is some B such that $B \in Cog$ and B weakly dominates A in the game restricted to Cog. Then there exists some strategy $C \in Cog$ such that $Pay_i(A, C) < Pay_i(B, C)$. Hence $Pay_i(A, C) = 1 - \epsilon$, which implies that $A \subset C$. But then A is not maximally cogent, which contradicts that it is a sustainable extension of AF. ($\Leftarrow$) Assume that A is not weakly dominated in the game restricted to Cog, and Cog is finite. Suppose by contradiction that $Pay_i(A, B) = 1 - \epsilon$ for some $B \in Cog$. Then, by Lemma 25 there exists some strategy $C \in Cog$ such that $Pay_i(C, B)$, i.e. C weakly dominates A, contradicting the hypothesis. Hence A always pays 1 against strategies in Cog, which means that it is maximally cogent, i.e. a sustainable extension of AF. □

Proposition 29. *An argument is in a sustainable strategy iff it can be defended in every* TPI*-*dispute*.

Proof. ($\Rightarrow$) Assume that A is a sustainable extension. Since A is conflict-free, CON cannot make an *eo ipso* move. Now, CON cannot block PRO through any argument β, because in that case $Pay(A, \{\beta\}) = (0, 1)$ (A would not be cogent), which contradicts the hypothesis. Hence, A can be defended in every TPI*-dispute.

($\Leftarrow$) Assume α is an argument that can be defended in every TPI*-dispute. For any optimal opposition of CON, assume $\alpha \in A$ where A is the strategy chosen by PRO. Assume that A is not conflict-free. Then CON could have made an *eo ipso* move. But this is not the case since α can be defended in each dispute. Therefore A is conflict-free. Even so, it may happen that $Pay(A, C) = (0, 1)$ for some strategy C that survives a IEWDS process. Then for some $\alpha_i \in A$ and $\beta \in C$, β attacks α_i but A does not attack β. But then β would be a winning argument, and α could not be defended in a dispute in which CON utters β. Contradiction. Therefore, A is tolerant. □

References

[1] L. Amgoud, Contribution a l'integration des préferences dans le raisonnement agumentatif, PhD thesis Université Paul Sabatier, Toulouse, July 1999.
[2] L. Amgoud, C. Cayrol, A reasoning model based on the production of acceptable arguments, AMAI Journal 34 (2002) 197–216.
[3] P. Baroni, M. Giacomin, Solving semantic problems with odd-length cycles in argumentation, in: Proc. of 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2003), Aalborg, Denmark, in: LNAI, vol. 2711, Springer-Verlag, 2003, pp. 440–451.
[4] P. Baroni, M. Giacomin, G. Guida, SCC-recursiveness: A general schema for argumentation semantics, Artificial Intelligence 165 (2) (2005) 187–259.
[5] P. Besnard, S. Doutre, Characterization of semantics for argument systems, in: Principles of Knowledge Representation and Reasoning (KR'04), Whistler, BC, Canada, AAAI Press, 2004, pp. 183–193.
[6] T. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, Journal of Logic and Computation 13 (3) (2003) 429–448.
[7] M. Caminada, L. Amgoud, On the evaluation of argumentation formalisms, Artificial Intelligence 171 (5–6) (2007) 286–310.
[8] C. Cayrol, C. Devred, M.-C. Lagasquie-Schiex, Handling controversial arguments in bipolar argumentation systems, in: Proc. of the 1st Conference on Computational Models of Argument (COMMA 2006), Liverpool, United Kingdom, 2006, pp. 261–272.

[9] S. Coste-Marquis, C. Devred, P. Marquis, Prudent semantics for argumentation frameworks, in: Proc. of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05), Hong-Kong, 2005, pp. 568–572.
[10] P.M. Dung, On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming, and n-person games, Artificial Intelligence 77 (1995) 321–357.
[11] S. Kaci, L. van der Torre, E. Weydert, Acyclic argumentation: Attack = conflict + preference, in: 17th European Conference on Artificial Intelligence, 2006, pp. 725–726.
[12] H.E. Kyburg, Probability and the Logic of Rational Belief, Wesleyan University Press, Middletown, CT, 1961.
[13] T. Kuhn, The Structure of Scientific Revolutions, University of Chicago Press, Chicago, 1962.
[14] H. Jakobovits, D. Vermeir, Robust semantics for argumentation frameworks, Journal of Logic and Computation 9 (2) (1999) 215–261.
[15] M. Osborne, A. Rubinstein, A Course in Game Theory, MIT Press, Cambridge, MA, 1994.
[16] J. Pollock, Nomic Probability and the Foundation of Induction, Oxford University Press, New York, 1990.
[17] J. Pollock, Self-defeating arguments, Minds and Machines 1 (1991) 367–392.
[18] J. Pollock, Justification and defeat, Artificial Intelligence 67 (1994) 377–407.
[19] J. Pollock, Cognitive Carpentry: A Blueprint for How to Build a Person, The MIT Press, 1995.
[20] D. Poole, On the comparison of theories: Preferring the most specific explanation, in: Proc. of the Ninth IJCAI, Los Altos, 1985, pp. 144–147.
[21] H. Prakken, Coherence and flexibility in dialogue games for argumentation, Journal of Logic and Computation 15 (2005) 1009–1040.
[22] G. Simari, R. Loui, A mathematical treatment of defeasible reasoning and its implementation, Artificial Intelligence 53 (1992) 125–157.
[23] G. Vreeswijk, Abstract argumentation systems, Artificial Intelligence 90 (1997) 225–279.
[24] G. Vreeswijk, H. Prakken, Credulous and sceptical argument games for preferred semantics, in: Proc. JELIA 2000, LNAI, vol. 1919, 2000, pp. 239–253.