Abstract Argumentation and the Rational Man

Timotheus Kampik
Juan Carlos Nieves
Department of Computing Science, Umeå University
90187 Umeå, Sweden

Abstract

Abstract argumentation has emerged as a method for non-monotonic reasoning that has gained tremendous traction in the symbolic artificial intelligence community. In the literature, the different approaches to abstract argumentation that were refined over the years are typically evaluated from a logics perspective; an analysis that is based on models of ideal, rational decision-making does not exist. In this paper, we close this gap by analyzing abstract argumentation from the perspective of the rational man paradigm in microeconomic theory. To assess under which conditions abstract argumentation-based choice functions can be considered economically rational, we define a new argumentation principle that ensures compliance with the rational man’s reference independence property, which stipulates that a rational agent’s preferences over two choice options should not be influenced by the absence or presence of additional options. We show that the argumentation semantics as proposed in Dung’s classical paper, as well as all of a range of other semantics we evaluate do not fulfill this newly created principle. Consequently, we investigate how structural properties of argumentation frameworks impact the reference independence principle, and propose a restriction to argumentation expansions that allows all of the evaluated semantics to fulfill the requirements for economically rational argumentation-based choice. For this purpose, we define the rational man’s expansion as a normal and non-cyclic expansion. Finally, we put reference independence into the context of preference-based argumentation and show that for this argumentation variant, which explicitly model preferences, the rational man’s expansion cannot ensure reference independence.

1. Introduction

Argumentation has been studied as a tool for persuasion and reasoning since antiquity. In recent decades, a large body of research emerged on argumentation as a method to instill intelligence, i.e. reasoning capabilities, into computing systems, in particular into autonomous agents. In this context, a popular theory for formal argumentation is so-called abstract argumentation, as initially developed by Dung (Dung, 1995), which is frequently used as the foundation of research on formal argumentation and hence spawned a variety of frameworks that extend the initial work (Baroni, Caminada, & Giacomin, 2018). Abstract argumentation is considered (see, e.g., Rahwan and Simari (Rahwan & Simari, 2009)) a step towards Leibniz’s vision of the unambiguous resolution of disagreements through mathematical means—the axiomization of argumentation:
If “controversies were to arise, there would be no more need of disputation; [it would suffice to say:] let us calculate” (Russell, 1938, p. 170).

A key aspect of abstract argumentation research is the definition and evaluation of different argumentation semantics. An argumentation semantics is a definition of how a graph of arguments—an argumentation framework—should be resolved, i.e. which arguments can be considered accepted and which cannot. Considering the variety of argumentation semantics (which determine how argumentation frameworks are to be resolved), as well as the range of advanced frameworks that aim to augment Dung’s basic definition of an argumentation framework, it is not clear how such a “controversy” is to be resolved in the sense of Leibniz and consequently what the result of a calculation process for argument resolution should be.

From the perspective economic theory, a classical model of choice is the notion of the rational man (see, e.g., Rubinstein (Rubinstein, 1998)). While the rational man as a sufficiently precise model of human decision-making has been debunked by a body of empirical research (Kahneman, 2003), it can still be considered relevant, for example for prescriptive modeling of an ideal agent that strictly optimizes according to clear preferences when determining its choices. Aligning abstract argumentation with the rational man paradigm requires, as this paper shows, the introduction of a new argumentation principle—an axiomatic requirement to an argumentation semantics (van der Torre & Vesic, 2017)–that Dung-style argumentation semantics typically do not fulfill. In contrast, applying argumentation semantics that do not guarantee rational behavior can lead to boundedly rational behavior that reflects human decision-making fallacies (see, e.g., Kahneman (Kahneman, 2003)).

In this paper, we use a novel formal approach to explore the intersection of abstract argumentation and (bounded) economic rationality and show that abstract argumentation can lead to economically not rational choice processes.

Let us provide an example that illustrates the problem this paper focusses on. An agent $A_1$ uses abstract argumentation for choosing a subset $A^*$ from a set $\{a, b\}$. For this, the agent constructs the attack relation $(a, b)$, i.e., $a$ attacks $b$. Given the set of attack relations $\{(a, b)\}$, the arguments $\{a, b\}$ are typically resolved as $\{a\}$. Let us refer to the constructed argumentation framework as $AF$. The agent’s choice of $\{a\}$ implies that the agent prefers choosing $\{a\}$ over all other possible options: $\{a\} \succeq \{a, b\}, \{a\} \succeq \{b\}, \{a\} \succeq \{\}$. Now, let us assume that a second agent $A_2$ consults $A_1$ to consider a third item $c$, i.e. the set to choose from is now $\{a, b, c\}$. $A_2$ also recommends to $A_1$ to add the attack relations $(b, c)$ and $(c, a)$. Given the attack relations $\{(a, b)(b, c), (c, a)\}$, many argumentation semantics resolve the arguments $\{a, b, c\}$ as $\{\}$. Let us refer to the constructed argumentation framework as $AF''$. If $A_1$ were to adopt the recommended attack relations, this implies that it now prefers $\{\}$.

---

1. While Caminada introduces rationality postulates for abstract argumentation semantics (Caminada, 2019), his work focuses on rationality from a logics perspective, i.e., on properties of argumentation semantics in regards to specific argumentation frameworks, without considering the expansions of any argumentation framework. Also, let us note that an application of the rational man’s argumentation principles we establish in this work to Amgoud’s and Cayrol’s preference-based argumentation is provided in Section 9.
over all elements \( \neq \{\} \) in \( 2^{\{a,b,c\}} \), which in turn implies \( \{\} \succeq \{a\} \). This is inconsistent with the preference \( \{a\} \succeq \{\} \) that \( \mathcal{A}_1 \) has given the set \( \{a,b\} \) (see Figure 1). In this case, \( \mathcal{A}_1 \) can be considered to behave \textit{economically irrationally}. The phenomenon that the presence of additional choice options changes existing preferences—in this case \( \{a\} \succeq \{\} \) to \( \{\} \succeq \{a\} \) is called \textit{reference dependence}. It is interesting to see that this change in preferences can be achieved without adding direct attack relations between arguments in \( AF \); \textit{i.e.} \( AF' \) can be referred to as a \textit{normal expansion} of \( AF \). Hence, \( \mathcal{A}_2 \) can attempt to deceive \( \mathcal{A}_1 \) into changing its preferences without noticing. From the perspective of economic rationality, the existence of the \textit{defeated} argument \( c \) should not change the preferences \( \mathcal{A}_1 \) has already established given the arguments \( \{a,b\} \). Indeed, such a change of preferences can be problematic in practice. For example, when an intelligent system needs to determine a set of actions, choosing \( \{a\} \) given \( \{a,b\} \), while choosing \( \{\} \) given \( \{a,b,c\} \) does typically not make sense, \textit{ceteris paribus}. Figure 1 depicts the argumentation graphs of \( AF \) and \( AF' \).

To address the problem of \textit{reference dependence} in abstract argumentation, this paper provides the following contributions:

1. Firstly, it characterizes the concept of reference independence in the context of argumentation semantics and argumentation framework expansions.

2. Secondly, it proofs that well-established argumentation semantics cannot guarantee reference independence for normal argumentation framework expansions.

3. Thirdly, it proofs that argumentation that is based on normal, non-cyclic expansions—which we refer to as \textit{rational man’s expansions}—guarantees reference independence.

4. Finally, it puts \textit{reference independence} and \textit{rational man’s expansions} in the context of preference-based argumentation, an argumentation approach that explicitly models agent preferences. This shows that reference dependence is also a problem for preference-based argumentation, which can, however, in this case not be solved with the \textit{rational man’s expansion} that this paper introduces.

The rest of this paper is organized as follows. Section 2 grounds this paper in the context of research traditions at the intersection of economics and artificial intelligence, whereas Section 3 provides an overview of abstract argumentation frameworks, semantics and principles, and of (bounded) rationality. Then, Section 4 presents argumentation frameworks.
in the context of choice functions for economic decision-making. Section 5 establishes the argumentation principles the rational man paradigm requires. Hereby, the new rational man’s reference independence principle is defined, based on which Section 6 shows that typically, argumentation semantics do not guarantee a rational choice process. To still allow for abstract argumentation in accordance with the rational man paradigm, a new argumentation framework expansion that helps guarantee economic rationality is defined in Section 7. Subsequently, Section 8 provides examples of how our concept of rational man’s argumentation expansions can be applied in different argumentation-based choice scenarios. Finally, Section 9 shows that the rational man’s expansion is also useful in the context of preference-based argumentation, before Section 10 concludes the paper by highlighting relevant future work.

2. Economic Rationality and Artificial Intelligence

Artificial intelligence frequently draws from concepts that have first been established in economic theory. For example, the notions of utility functions and preferences that are central to many models and algorithms of autonomous agents and multi-agent systems stem from ideas of the philosophers and economists Jeremy Bentham and John Stuart Mill (Sen, 1991). Von Neumann’s and Morgenstern’s ground-breaking game theoretical work Theory of Games and Economic Behavior (Von Neumann, Morgenstern, & Kuhn, 2007) influenced generations of both economics and artificial intelligence researchers. More recently, the works of Daniel Kahneman and Richard Thaler, who both have received the Nobel Memorial Prize in Economic Sciences for their research on behavioral economics, have inspired new work on and discussions about ethics and responsibility in artificial intelligence (Dignum, Baldoni, Baroglio, Caon, Chatila, Dennis, Génova, Haim, Klień, Lopez-Sanchez, Micalizio, Pavón, Slavkovik, Smakman, van Steenbergen, Tedeschi, van der Torre, Villata, & de Wildt, 2018).

A key concept at the intersection of economics and artificial intelligence is the notion of economic rationality. In the same way that a rational economic decision-maker—traditionally a human agent—is ideally expected to act according to clear and consistent preferences (Rubinsteint, 1998, p. 7 et sqq.), artificially intelligent agents should “maximize [their] performance measure” as stated in the definition of a rational agent by Russel and Norvig (Russell & Norvig, 2016, p. 37). The concept of economical rationality is reflected in many ground-breaking works on economics, such as the Nash equilibrium (Nash, 1950), and Tversky’s and Kahneman’s empirical work on the limits of human agents to act rationally in the economic sense (Kahneman, 2003).

An important property in the context of economic rationality is reference independence: a decision-maker’s preferences over a set of items \( S \) should not be affected by the presence, or absence, of additional items \( S^+ \). The negation of reference independence—reference dependence—as established as a theory by Tversky and Kahneman (Tversky, Slovic, & Kahneman, 1990) and empirically validated as a phenomenon in human decision-making by, e.g., Bateman et al. (Bateman, Munro, Rhodes, Starmer, & Sugden, 1997), stipulates that an agent might change their initial preference \( a \succeq b \) out of the set of possible choices \( A \), with \( a, b \in A \) to \( b \succeq a \), depending on reference points that are either added to the choice set itself or provided as additional context, but that do not impact the value or quality of
either $a$ or $b$. As a real-world example, let us summarize a study by Doyle et al. (Doyle, O’Connor, Reynolds, & Bottomley, 1999). In a grocery store, two brands of baked beans $x$ and $y$ are sold. $x$ and $y$ are sold in cans of the same size. Although $x$ is cheaper than $y$, only 19% of bean sales are of brand $x$. By adding a new option $x'$—a smaller can of beans by brand $x$ that is sold at the same price as the original can $x$—the share of $x$ increases to 33% of total sales (while sales of $x'$ are negligible).

From the body of research we summarize above it is clear that human decision-making is often not reference independent. However, reference independence can typically be considered a desired property of the decision-making process of artificially intelligent agents, as we have demonstrated above, in the example that is visualized by Figure 1. Yet, when evaluating artificially intelligence systems, economic rationality in general, and reference dependence in particular, is typically not considered. Instead, research results are commonly evaluated based on traditional performance criteria like accuracy and computational complexity. In this context, a group of well-established artificial intelligence researchers advocates for a paradigm shift in the evaluation of “machine behavior” (Rahwan, Cebrian, Obradovich, Bongard, Bonnefon, Breazeal, Crandall, Christakis, ..., & Wellman, 2019) that can be considered a continuation of the cross-disciplinary work of Herbert Simon, who initially coined the term bounded rationality (Simon, 1955). In the spirit of Simon, but also of Kahneman, who uses empirical methods to systematically specify the boundaries of human rationality, this paper presents an application of the study of economic rationality to the area of formal argumentation. As we study—in contrast to Kahneman—formally specified frameworks and not humans, we are able to use formal methods in our evaluations. Thereby, our focus lies on an artificial intelligence method that can, as described above, be traced to the ideas of Leibniz, and that should, as a tool to objectively resolve disagreements, ideally reflect the principles of rational economic decision-making.

### 3. Theoretical Background

This section introduces the foundations upon which this paper builds: the rational man paradigm in economic theory and abstract argumentation frameworks, in particular argumentation semantics and principles.

#### 3.1 The Rational Economic Man Paradigm

As a prerequisite, we introduce a definition of a partially ordered set (Davey & Priestly, 2002).

**Definition 1.** Let $Q$ be a set. An order (or partial order) on $Q$ is a binary relation $\succeq$ on $Q$ such that, for all $x, y, z \in Q$:

1. $x \succeq x$ (reflexivity);
2. $x \succeq y$ and $y \succeq x$ imply $x = y$ (antisymmetry);
3. $x \succeq y$ and $y \succeq z$ imply $x \succeq z$ (transitivity).

We refer to a set $Q$ that is equipped with an order relation $\succeq$ as an ordered set (or partial ordered set) and denote it by $(Q, \succeq)$. In economic theory, the model of a rational deci-
Figure 2: Example: a graphical representation of a choice set, in which the following preference relations of the rational man hold: \( \forall A' \in \{ \{ \text{coffee} \}, \{ \text{snack} \}, \{ \} \} : \{ \text{coffee, snack} \} \succeq A' \) (\( a \succeq b \) is visualized by \( a \rightarrow b \)). Note that the preference graph is not an argumentation framework, as indicated by the different style of arrows.

The choice maker—the \textit{rational economic man} paradigm\(^2\) can be described as follows (based on a definition by Rubinstein (Rubinstein, 1998, p. 7 et sqq.))\(^3\).

**Definition 2** (Rational Economic Man). \textit{Given a set of choice options} \( A \), the \textit{rational economic man} determines his choice of elements \( A^* \subseteq A \) according to his preference order (partial order) \( \succeq \) so that \( \forall A' \in 2^A : A^* \succeq A' \); i.e., \( A^* \) is the rational man's preferred option when compared to all possible alternatives.

For example, given the choice set \( \{ \text{coffee, snack} \} \), the rational man’s preferred option could be \( \{ \text{coffee, snack} \} \), given he chooses this element from the set \( 2^{\{ \text{coffee, snack} \}} \). The choice implies \( \forall A' \in \{ \{ \text{coffee} \}, \{ \text{snack} \}, \{ \} \} : \{ \text{coffee, snack} \} \succeq A' \), as depicted in Figure 2.

In the context of this paper, an important property of the rational man’s choice process is \textit{reference independence}.

**Definition 3** (Reference Independence). \textit{Given a set of choice options} \( A \) and any two possible subsets of choice options \( A_1 \subseteq A \) and \( A_2 \subseteq A \), with \( A_1 \subseteq A_2 \) and the rational man’s choices \( A_1^* \subseteq A_1 \) (based on the choice options \( A_1 \)) and \( A_2^* \subseteq A_2 \), if \( A_2^* \subseteq A_1 \), then \( A_2^* = A_1^* \)\(^4\).

Colloquially expressed, the rational man’s choice outcome is not affected if options that the agent does not prefer over the “previously” existing options are added to the set of potential choices.

---

2. In this paper, we often use the shortened term \textit{rational man}.
3. However, in contrast to the model used by Rubinstein, we use a choice model that allows for choosing any subset \( A^* \subseteq A \) of the set of choice options \( A \) instead of exactly one option (\( x^* \in A \))
4. A proof of this property for the “chose one” variant of the model is provided by Rubinstein (Rubinstein, 1998, p. 11)
Example 1. If a decision-maker, given the set \{\text{tee, cookie}\} chooses \{\text{tee}\} and given the set \{\text{tee, cookie, coffee}\} chooses \{\text{tee, cookie}\} the choice is not rational; the presence of the “new” irrelevant alternatives \{\text{coffee}\}, \{\text{tee, coffee}\}, \{\text{cookie, coffee}\}, and \{\text{tee, cookie, coffee}\} causes the preference relation \{\text{tee}\} \succeq \{\text{tee, cookie}\} to change to \{\text{tee, cookie}\} \succeq \{\text{tee}\}.

The reference independence property implies that the expansion of a choice set \(A_1\) to \(A_2 \supset A_1\) does not add new knowledge about the elements in \(A_1\) that affects the decision-maker’s preferences over elements in \(2^A_1\). Yet, in reality, human decision-makers can use the addition of options to a choice set as a way to infer new information about the quality of the original choice options, as for example shown by Doyle et al. (Doyle et al., 1999). Still, in the context of computational argumentation-based choice, it can be useful to be able to distinguish between changes in choice options and outcomes that comply with the reference independence property and those that do not. For example, in a consultation scenario, a consulted agent may want to check if the consulting agent is providing a proposal that is seemingly barely adding new choice options, but covertly also altering the preferences over existing choice options. Such a scenario is presented in greater detail in Section 8.

3.2 Abstract Argumentation

To allow for a concise overview of the relevant foundations of abstract argumentation, we introduce a formal definition of the basic structure of an argumentation framework (AF), provide the definitions of well-established argumentation semantics, and explain the notion of argumentation expansions and principles.

Definition 4 (Argumentation Framework). (Dung, 1995) An argumentation framework is a pair \(AF := (AR, \text{Attacks})\), where \(AR\) is a finite set of arguments, and \(\text{Attacks}\) is a binary relation on \(AR\), i.e. \(\text{Attacks} \subseteq AR \times AR\).

In the context of an argumentation framework, \(a\) attacks \(b\) means that \(a\) attacks \((a, b)\) holds. Similarly, \(a\) set \(S\) attacks \(b\) if \(b\) is attacked by an argument in \(S\), and \(c\) defends \(b\) if \(a\) attacks \(b\) and \(c\) attacks \(a\).

To select coherent sets of arguments from an argumentation framework, Dung introduces the notion of conflict-free, acceptable, and admissible sets.

Definition 5 (Conflict-free, Acceptable, and Admissible Arguments). (Dung, 1995)

- **Conflict-free.** A set of arguments \(S\) is conflict-free iff there are no arguments \(a, b\) in \(S\) such that \(a\) attacks \(b\).

- **Acceptable.** Given an argumentation framework \(AR\), an argument \(a \in AR\) is acceptable w.r.t. an argument set \(S\) iff for each argument \(b \in AR\): if \(b\) attacks \(a\), then \(b\) is attacked by \(S\).

- **Admissible.** A conflict-free set of arguments \(S\) is admissible iff each argument in \(S\) is acceptable with regards to \(S\).
Based on these concepts, Dung (and others) define different *semantics* for resolving argumentation frameworks.

**Definition 6 (Argumentation Semantics).** (Dung, 1995) An argumentation semantics $\sigma$ is a function that takes an argumentation framework $AF = (AR, \text{attacks})$ and returns a set of sets of arguments denoted by $\sigma(AF)$.

$$\sigma : AF \rightarrow 2^{2^{AR}},$$

with $AF$ denoting the set of all possible argumentation frameworks. Each set of $\sigma(AF)$ is called a $\sigma$-extension.

Below, we provide an overview of semantics/extensions that are commonly applied and analyzed by the formal argumentation community. In the context of this paper, we use the colloquial umbrella term *well-established semantics* to refer to the semantics introduced in Definition 7\(^5\).

**Definition 7 (Argumentation Extensions).** (Dung, 1995; Dung, Mancarella, & Toni, 2007; Caminada, Carnielli, & Dunne, 2012; Caminada, 2007; Verheij, 1996; Baroni et al., 2018) Given an argumentation framework $AF = (AR, \text{Attacks})$, an admissible argument set $S \subseteq AR$ is a:

- **stable extension** ($\sigma_{\text{stable}}$) of $AF$ iff $S$ attacks each argument that does not belong to $S$.
- **preferred extension** ($\sigma_{\text{preferred}}$) of $AF$ iff $S$ is a maximal (w.r.t. set inclusion) admissible set of $AF$.
- **complete extension** ($\sigma_{\text{complete}}$) of $AF$ iff each argument that is acceptable w.r.t. $S$ belongs to $S$.
- **grounded extension** ($\sigma_{\text{grounded}}$) of $AF$ iff $S$ attacks each argument that does not belong to $S$.
- **ideal extension** ($\sigma_{\text{ideal}}$) of $AF$ iff $S$ is contained in every preferred extension of $AF$.
- **semi-stable extension** ($\sigma_{\text{semi-stable}}$) of $AF$ iff $S$ is a complete extension where $S \cup S^+$ is maximal (w.r.t. set inclusion), with $S^+ := \{b | \exists a \in AR : (a, b) \in \text{Attacks}\}$.
- **eager extension** ($\sigma_{\text{eager}}$) of $AF$ iff $S$ is the greatest (w.r.t. set inclusion) admissible set that is a subset of each semi-stable extension.

Given an argumentation framework $AF = (AR, \text{Attacks})$, a conflict-free argument set $S \subseteq AR$ is a **stage extension** ($\sigma_{\text{stage}}$) of $AF$, iff $S$ is conflict-free and $S \cup S^+$ is maximal (w.r.t. set inclusion), with $S^+ := \{b | \exists a \in AR : (a, b) \in \text{Attacks}\}$.

---

\(^5\) This does not mean that in general, other semantics cannot be considered as *well-established*. 

8
To introduce the definitions of \textit{stage2} and \textit{CF2} semantics, let us first define the notion of \textit{strongly connected components} as follows. “Strongly connected components (SCCS) provide a unique partition of a directed graph into disjoint parts where all nodes are mutually reachable” (Baroni et al., 2018). In the context of an argumentation framework \textit{AF}, the set of strongly connected components of \textit{AF} is denoted by \textit{SCCS}_{\textit{AF}}. For example, the graph of the argumentation framework ((\{a, b, c\}, \{(a, b), (b, a), (b, c)\}) has the SCCS \{a, b\} and \{c\}.

**Definition 8 (CF2 Extensions).** (Baroni et al., 2018) Given an argumentation framework \textit{AF} = \textit{(AR, Attacks)}, a set of arguments \textit{S} \subseteq \textit{AR} is an extension of CF2 semantics (\textit{\sigma}_{\textit{CF2}}) iff:

- if \(|\textit{SCCS}_{\textit{AF}}| = 1\), \textit{S} is a maximal conflict-free extension of \textit{AF};
- otherwise, \forall \textit{Args} \in \textit{SCCS}_{\textit{AF}} (\textit{S} \cap \textit{Args}) \in \textit{\sigma}_{\textit{CF2}} (\textit{AF} \downarrow_{\textit{UP}_{\textit{AF}} (\textit{Args}, \textit{S})}) , with \textit{AF} \downarrow_{\textit{UP}_{\textit{AF}} (\textit{Args}, \textit{S})} = (\textit{UP}, \textit{S} \cap \textit{UP} \times \textit{UP}) , the restriction of \textit{AF} to \textit{UP} (Baroni, Giacomin, & Guida, 2005) and for any \textit{S}, \textit{Args} \subseteq \textit{AR}, \textit{UP}_{\textit{AF}} (\textit{Args}, \textit{S}) = \{a \in \textit{Args} \mid \exists b \in \textit{S} \text{ Args} : (b, a) \in \textit{Attacks}\}.

**Definition 9 (Stage2 Extensions).** (Baroni et al., 2018) Given an argumentation framework \textit{AF} = \textit{(AR, Attacks)}, a set of arguments \textit{S} \subseteq \textit{AR} is an extensions of stage2 semantics (\textit{\sigma}_{\textit{stage2}}), iff:

- if \(|\textit{SCCS}_{\textit{AF}}| = 1\), \textit{S} is a stage extension of \textit{AF};
- otherwise \forall \textit{Args} \in \textit{SCCS}_{\textit{AF}} (\textit{S} \cap \textit{Args}) \in \textit{\sigma}_{\textit{stage2}} (\textit{AF} \downarrow_{\textit{UP}_{\textit{AF}} (\textit{Args}, \textit{S})}) otherwise, with \textit{AF} \downarrow_{\textit{UP}_{\textit{AF}} (\textit{Args}, \textit{S})} = (\textit{UP}, \textit{S} \cap \textit{UP} \times \textit{UP}) , the restriction of \textit{AF} to \textit{UP} (Baroni et al., 2005) and for any \textit{S}, \textit{Args} \subseteq \textit{AR}, \textit{UP}_{\textit{AF}} (\textit{Args}, \textit{S}) = \{a \in \textit{Args} \mid \exists b \in \textit{S} \text{ Args} : (b, a) \in \textit{Attacks}\}.

In words, CF2 and stage2 semantics decomposition an argumentation framework into SCCS and recursively resolve the framework component by component, starting with the SCCS that are not attacked by any other SCC. Note that a detailed explanation of CF2 and stage2 semantics is beyond the scope of the paper.

Finally, we provide the established definitions for credulous and skeptical acceptance of arguments (Doutre & Mengin, 2004), given a specific argumentation semantics and in addition introduce the concept of lenient acceptance to allow for more concise formal work.

**Definition 10 (Credulous, Skeptical, and Lenient Acceptance).** (Doutre & Mengin, 2004)

- **Credulous Acceptance.** Given an argumentation framework \textit{AF} = \textit{(AR, Attacks)} and an argumentation semantics \textit{\sigma_x}, an argument \textit{a} \in \textit{AR} is credulously accepted w.r.t. \textit{\sigma_x} iff it is contained in at least one of the extensions of \textit{\sigma_x}(\textit{AF}).
• **Skeptical Acceptance.** Given an argumentation framework \( AF = (AR, \text{Attacks}) \) and an argumentation semantics \( \sigma_x \), an argument \( a \in AR \) is skeptically accepted iff it is contained in all of the extensions of \( \sigma_x(AF) \). The skeptical extension of an argumentation framework \( AF \) is the intersection of all extensions \( \sigma_x(AF) \) (a.i., \( \sigma_x^S(AF) := \bigcap_{E \in \sigma_x(AF)} E \)).

• **Lenient Acceptance.** Given an argumentation framework \( AF = (AR, \text{Attacks}) \) and an argumentation semantics \( \sigma_x \), the lenient extension of \( AF \) is the union of all extensions \( \sigma_x(AF) \) (a.i., \( \sigma_x^L(AF) := \bigcup_{E \in \sigma_x(AF)} E \)).

Let us note that the extensions returned by lenient semantics are not necessarily conflict-free.

Given the variety of argumentation semantics that have been established throughout the years, it can be challenging to assess which semantics are suitable for a specific application. Consequently, *argumentation principles* have been established; evaluating the compliance with one or multiple principles can guide the assessment of argumentation semantics in the context of a specific use case. For example, a principle can specify that an argumentation semantics should prescribe exactly one extension for any argumentation framework. An overview of argumentation principles is provided by van der Torre and Vesic (van der Torre & Vesic, 2017).

### 3.3 Argumentation Expansions

To describe a specific type of relation between two argumentation frameworks, the notion of an argumentation framework *expansion* has been introduced by Baumann and Brewka (Baumann & Brewka, 2010). A concise introduction and overview of expansions is presented by Baumann and Woltran (Baumann & Woltran, 2016). The general concept of an argumentation framework expansion can be defined as follows.

**Definition 11 (Argumentation Framework Expansion).** An argumentation framework \( AF' \) is an expansion of an argumentation framework \( AF = (AR, \text{Attacks}) \) \( (AF \preceq_E AF') \) iff \( AF' = (AR \cup AR', \text{Attacks} \cup \text{Attacks'}) \), where \( AR' \cap AR = \text{Attacks'} \cap \text{Attacks} = \emptyset \).

Several expansion types have been defined in the literature. In the context of this paper, the notion of a *normal* expansion is of relevance.

**Definition 12 (Normal Expansion).** An expansion \( AF' \) of an argumentation framework \( AF \) is normal \( (AF \preceq_N AF') \) iff \( \forall a, b : ((a, b) \in \text{Attacks'}, a \in AR' \lor b \in AR') \).

In words, a normal expansion of an argumentation framework adds additional arguments to the framework that can attack and be attacked by any other argument but does neither remove existing arguments nor change attack relations (neither add nor remove) between existing arguments.

### 4. Rational Argumentation-based Choice Function

In this section, we define the concept of *economically rational argumentation-based choice functions* and, in this context, provide a brief overview of the difference between the rational man paradigm’s *reference independence* requirement and monotonicity.
4.1 Rational Argumentation-based Choice

To build the foundation for exploring the intersection of abstract argumentation semantics and (bounded) rationality, we introduce the notion of an argumentation-based choice function.

**Definition 13 (Argumentation-based Choice Function).** The argumentation-based choice function $g \circ \sigma$ of an agent is the function composition between a function $g$ and an argumentation semantics $\sigma$ that takes an argumentation framework as its input and returns a set of decision outcomes $AR^* \subseteq AR$:

$$g \circ \sigma : AF \rightarrow 2^{AR}$$

To determine the function’s input, an economically rational agent needs to construct an argumentation framework $AF$ that consists of the propositional atoms—called arguments—$AR$ and a set of binary attack relations between those arguments $Attacks \subseteq AR \times AR$. The framework is resolved by an argumentation semantics $\sigma$; different argumentation semantics can be used. In the context of rational economic choice, the mapping from a semantics’ output (a set of sets) to the choice function’s output requires that the semantics returns always exactly one extension. If this is the case, the choice function can be defined in different ways, for instance:

1. $g^\cup \circ \sigma(AF) = \bigcup_{E \in \sigma(AF)} E$. If a semantics $\sigma$ returns more than one extension, the choice function $g^\cup \circ \sigma(AF) = \bigcup_{E \in \sigma(AF)} E$ returns the union of the set of extensions returned by $\sigma$. We call such a function a lenient choice function.

2. $g^\cap \circ \sigma(AF) = \bigcap_{E \in \sigma(AF)} E$. If a semantics $\sigma$ returns more than one extension the choice function $g^\cap \circ \sigma(AF) = \bigcap_{E \in \sigma(AF)} E$ returns the intersection of the set of extensions returned by $\sigma$. We call such a function a skeptical choice function.

However, for the purpose of evaluating argumentation semantics, it makes sense to define the aggregation of extensions as part of the semantics itself, for example as follows:

1. $g^i \circ \sigma^\cup_{\text{stage}}(AF)$ returns the lenient stage extension of $AF$.

2. $g^i \circ \sigma^\cap_{\text{stage}}(AF)$ returns the skeptical stage extension of $AF$.

3. $g^i \circ \sigma^\cap_{\text{grounded}}(AF)$ and $g^i \circ \sigma^\cup_{\text{grounded}}(AF)$ return the grounded extension of $AF$. There is always exactly one grounded extension, from which it follows that $g^i \circ \sigma^\cap_{\text{grounded}}(AF) = g^i \circ \sigma^\cup_{\text{grounded}}(AF)$. The same applies to ideal and eager semantics.

For this, we characterize $g^i$ as the identity function:

$$g^i \circ \sigma(AF) = \sigma(AF)$$

It is important to note that argumentation-based choice does not necessarily imply choice from a set of goods or indeed any type of scenario that is typical for economic decision-making examples, but can cover any decision-making process, for example the selection of epistemic arguments from an argumentation framework that an agent will consider as valid.
Example 2. As an example, let us consider the choice process of a family that wants to purchase a car. The local car dealer has a limited selection of cars \( AR = \{ \text{volvo, polo, porsche} \} \). As the family prefers a safe car with reasonable purchase and maintenance costs, the attacks can be specified as \( \text{Attacks} = \{ (\text{volvo, polo}), (\text{volvo, porsche}), (\text{polo, porsche}) \} \), which gives us the following argumentation framework \( AF \):

\[
AF = ( \\
   \{ \text{volvo, polo, porsche} \}, \\
   \{ (\text{volvo, polo}), (\text{volvo, porsche}), (\text{polo, porsche}) \})
\]

Solving the framework, using for example complete semantics, returns \( \{ \{ \text{volvo} \} \} \). Consequently, given skeptical complete semantics \( \sigma_c^\prime \), the argumentation-based choice function \( g^\prime \circ \sigma_c^\prime (AF) \) returns \( \{ \text{volvo} \} \), which implies (given an economically rational family) \( \forall A \in 2^{\{\text{volvo, polo, porsche}\}} : \{ \text{volvo} \} \succeq A \); even if presented with a choice from a larger set \( AR' \supset AR \) these preferences need to hold.

Example 3. As a slightly more complex example, let us consider the choice process of a person who considers buying a beverage in a café. The café only offers tea and coffee: \( AR = \{ \text{tea, coffee} \} \). The person prefers the caffeine kick of coffee, but also acknowledges the health advantages of tea: \( \text{Attacks} = \{ (\text{coffee, tea}), (\text{tea, coffee}) \} \). Complete semantics \( \sigma_c \) resolve the resulting framework \( AF = (AR, \text{Attacks}) \) as \( \{ \{ \text{tea} \}, \{ \text{coffee} \}, \{ \} \} \). An argumentation-based choice function \( g^\prime \circ \sigma_c^\prime \) based on lenient complete semantics \( \sigma_c^\prime \) flattens this set of sets by determining the union of the complete extensions \( \{ \text{coffee, tea} \} \), implying \( \forall A \in 2^{\{\text{tea, coffee}\}} : \{ \text{coffee, tea} \} \succeq A \). In contrast, \( g^\prime \circ \sigma_c^\prime \), which is based on skeptical complete semantics, returns \( \{ \} \), implying \( \forall A \in 2^{\{\text{tea, coffee}\}} : \{ \} \succeq A \).

Let us now put argumentation-based choice functions in the context of economic rationality. By considering Definition 2, we define the clear preferences property for argumentation-based choice.

Definition 14 (Clear Preferences for Argumentation-based Choice).
Let \( AF = (AR, \text{Attacks}) \) be an argumentation framework and \( g \circ \sigma \) an argumentation-based choice function. An argumentation-based choice \( A^* = g \circ \sigma(AF) \) is economically rational iff it holds that \( \forall A \in 2^{AR}, A^* \succeq A \).

In words, the agent’s choice \( A^* \) from \( 2^{AR} \) means that the agent prefers \( A^* \) over any of the other sets in \( 2^{AR} \).

The clear preferences principle is obvious when evaluating a one-off argumentation-based choice. However, when considering an argumentation process or dialog, during which arguments and attacks are added to an argumentation framework over time, it is clear that we need to consider the reference independence property as introduced in Definition 3. We prove that, analogous to the economic principles, reference dependence is implied by the clear preferences property in the context of argumentation-based choice.

Proposition 1 (Reference Independence for Argumentation-based Choice).
Let \( AF = (AR, \text{Attacks}) \) and \( AF' = (AR', \text{Attacks}') \) be two argumentation frameworks for which it holds that \( AR \subseteq AR' \) and let \( g \circ \sigma \) be an argumentation-based choice function. An economically rational argumentation-based choice \( A^* = g \circ \sigma(AF) \) implies that if \( g \circ \sigma(AR') \subseteq AR \), then \( g \circ \sigma(AF') = g \circ \sigma(AF) \).
Proof. By contradiction, we prove that \( g \circ \sigma(AF) = g \circ \sigma(AF') \) given that:

- \( AF = (AR, Attacks) \),
- \( AF' = (AR', Attacks') \),
- \( AF \subseteq AF' \),
- \( g \circ \sigma(AF') \subseteq AR \).

Let us suppose that \( g \circ \sigma(AF) \neq g \circ \sigma(AF') \). \( \implies g \circ \sigma(AF) \nsubseteq g \circ \sigma(AF') \nsubseteq g \circ \sigma(AF) \).

1. If \( g \circ \sigma(AF) \nsubseteq g \circ \sigma(AF') \): \( \exists Arg \in g \circ \sigma(AF) \land Arg \nsubseteq g \circ \sigma(AF') \).
   \( \implies i) g \circ \sigma(AF) \rightarrow \exists A^* \in 2^{AR} \land \forall A \in 2^{AR} : A^* \succeq A \land Arg \in A^* ,
   ii) g \circ \sigma(AF') \rightarrow \exists A'^* \in 2^{AR} \land \forall A \in 2^{AR} : A'^* \succeq A \land Arg \nsubseteq A'^* .
   \implies i) \text{ contradicts ii).} \)

2. If \( g \circ \sigma(AF') \nsubseteq g \circ \sigma(AF) \): \( \exists Arg \in g \circ \sigma(AF') \land Arg \nsubseteq g \circ \sigma(AF) \).
   \( \implies i) g \circ \sigma(AF') \rightarrow \exists A'^* \in 2^{AR} \land \forall A \in 2^{AR} : A'^* \succeq A \land Arg \nsubseteq A'^* ,
   ii) g \circ \sigma(AF) \rightarrow \exists A^* \in 2^{AR} \land \forall A \in 2^{AR} : A^* \succeq A \land Arg \nsubseteq A^* .
   \implies i) \text{ contradicts ii).} \)

In words, given a choice \( A^* \) from \( 2^{AR} \) that implies \( \forall A \in 2^{AR} : A^* \succeq A \), no choice \( A' \) from \( 2^{AR'} \) with \( AR' \supseteq AR \) should change the preferences implied by \( A^* \).

### 4.2 Rationality and Monotonicity

Considering that abstract argumentation is a method for non-monotonic reasoning, and given that the reference independence property may seem–at first glance–to imply monotonicity, let us provide an intuition for distinguishing between the two concepts in the context of argumentation-based choice. Given a set to choose from \( AR \) and a choice \( A^* \) as a subset of \( AR \), the choice implies \( \forall A \in 2^{AR} : A^* \succeq A \). The reference independence property requires that for all \( AR' \supseteq AR \) the choice \( A'^* \supseteq AR \) complies with the preference relation implied by \( A^* \); i.e. \( A'^* = A^* \lor A'^* \nsubseteq AR \). In contrast, monotonicity requires that if the choice process determines a set \( A^* \subseteq AR \), given the possible choices \( 2^{AR} \), the choice \( A^* \) needs to be a subset of the choice set \( A'^* \subseteq AR' (AR' \supseteq AR) \), given the possible choices \( 2^{AR'} \).

For example, given the sets \( AR = \{a_1, a_2\} \) and \( AR' = \{a_1, a_2, a_3\} \), the reference independence property is maintained, given the following choice result of an agent \( i \):

- Given \( AR \), \( i \) chooses \( \{a_1\} \), i.e. \( i \) prefers \( \{a_1\} \) over all other possible choices (\( \{a_1\} \succeq \{a_2\}, \{a_1\} \succeq \{a_1, a_2\} \), \( \{a_1\} \succeq \{\} \)).
- Given \( AR' \), \( i \) chooses \( \{a_3\} \), i.e., \( i \) prefers \( \{a_3\} \) over all other possible outcomes (\( \forall A \in 2^{\{a_1, a_2, a_3\}} : \{a_3\} \succeq A \)).
Because the agent retracts its decision that \( \{a_1\} \) is in the chosen set, its choice process is non-monotonic. However, only if the agent were to retract its preferences, e.g., by deciding that, given \( AR' \), \( \{a_2\} \succeq \{a_1\} \), the reference independence property of the rational man paradigm would be infringed.

5. Argumentation Principles for the Rational Man

From the two rational man properties that we established in the context of argumentation-based choice, we derive two principles that an argumentation semantics \( \sigma \) needs to fulfill to guarantee rational argumentation-based choice, given a choice function \( g \circ \sigma \). The first principle is necessary to fulfill the rational man’s clear preferences property.

**Definition 15** (Clear Preferences Principle). An argumentation semantics \( \sigma \) fulfills the clear preferences principle, iff for any argumentation framework \( AF = (AR, aAtacks) \) the following applies:

\[
\forall AF \in \mathcal{AF} : |g \circ \sigma(AF)| = 1
\]

It is clear that all of the argumentation semantics introduced by Definitions 7, 8, and 9 with the exception of stable semantics, which—as shown by Dung—do not return any extension for some argumentation frameworks (Dung, 1995), fulfill the clear preferences principle, given a lenient choice function \( g \cup \circ \sigma \) or a skeptical choice function \( g \cap \circ \sigma \).

**Example 4.** For example, given the argumentation framework \( AF = (\{a,b\}, \{(a,a)\}) \), a skeptical choice function based on stable semantics \( g \cap \circ \sigma_{stable} \) returns no result, which implies the decision-maker does not have clear preference \( A^* \), with \( \forall A \in 2^{\{a,b\}} : A^* \succeq A \). In contrast skeptical choice functions based on other semantics could return \( \{b\} \), which implies \( A^* = \{b\} \succeq \{a,b\}, \{b\} \succeq \{a\}, \) and \( \{b\} \succeq \{\} \).

Note that an overview of semantics that fulfill the clear preferences principle without the need of an aggregating choice function \( g^\cup \circ \sigma \) or \( g^\cap \circ \sigma \) (by returning always exactly one extension) is provided in the Appendix.

The second principle is necessary to fulfill the rational man’s reference independence property.

**Definition 16** (Reference Independence Principle). An argumentation semantics \( \sigma \) fulfills the reference independence principle iff for any two argumentation frameworks \( AF = (AR, aAttacks) \) and \( AF' = (AR', aAttacks') \), with \( AF \preceq_N AF' \) the following applies, given an argumentation-based choice function \( g \circ \sigma \):

\[
g \circ \sigma(AF') \subseteq AR \implies g \circ \sigma(AF') = g \circ \sigma(AF)
\]

To illustrate the principle, let us introduce an example.

**Example 5.** We have the following argumentation frameworks:

1. \( AF_1 = (AR_1, aAttacks_1) = (\{a,b\}, \{(a,b)\}) \)
2. \( AF_2 = (AR_2, aAttacks_2) = (\{a,b,c\}, \{(a,b), (b,c), (c,a)\}) \)

Note that \( AF_1 \preceq_N AF_2 \). Now, we apply complete semantics \( \sigma_{complete} \) to both frameworks.

---

6. In this case, the aggregation as skeptical (\( \cap \)) returns the same results as a lenient aggregation (\( \cup \)). Also, many other semantics return the same set of arguments in this case.
Abstract Argumentation and the Rational Man

1. $g^i \circ \sigma^\cap_{\text{complete}}(AF_1) = \{a\}$, which implies $\forall \text{Args} \in 2^{\{a, b\}} : \{a\} \succeq \text{Args}$.

2. $g^i \circ \sigma^\cap_{\text{complete}}(AF_2) = \{\}$, which implies $\forall \text{Args} \in 2^{\{a, b, c\}} : \{\} \succeq \text{Args}$.

$g^i \circ \sigma^\cap_{\text{complete}}(AF') \subset \text{AR}$ and $g^i \circ \sigma^\cap_{\text{complete}}(AF') \neq g^i \circ \sigma^\cap_{\text{complete}}(AF)$; i.e., the choices of $\{a\}$ given $\{a, b\}$ and $\{\}$ given $\{a, b, c\}$ are not rational, because the preference orders they imply are inconsistent: $g^i \circ \sigma^\cap_{\text{complete}}(AF_1) \rightarrow \{a\} \succeq \{\}$, whereas $g^i \circ \sigma^\cap_{\text{complete}}(AF_2) \rightarrow \{\} \succeq \{a\}$.

Hence, it is clear that the argumentation semantics do not fulfill the reference independence principle.

Note that the principle applies by definition only to normal expansions of a framework $(AF \preceq_N AF')$. We introduced this restriction because it is obvious that any expansion $AF'$ of $AF = (AR, \text{Attacks})$ with $AF \not\preceq_N AF'$ that adds attack relations between arguments is revising the assumptions about framework $AF$ and hence is clearly only suitable for economically rational choice if a belief revision w.r.t. $AF$ takes place. However, one could define the additional principle strict reference independence that does not assume $AF \preceq_N AF'$, which then requires the introduction of $AF \preceq_N AF'$ as a necessary condition to enable rational choice. In any case, this distinction has little effect on the approach taken in this paper.

6. Reference Dependence in Abstract Argumentation

As shown above, most argumentation semantics can be considered compliant with the rational man’s clear preferences property when only considering the resolution of a specific argumentation framework, at least if they are skeptically $(\sigma^\cap(AF))$ or leniently $(\sigma^\cup(AF))$ aggregated. However, we can provide simple examples that proof non-compliance with the reference independence principle for complete, preferred, semi-stable, stage, stage2, and CF2 semantics (lenient and skeptical), as well as grounded, ideal and eager semantics; i.e., these argumentation semantics do not guarantee clear preferences in the context of dynamic argumentation processes or dialogues. These observations are formalized by the following propositions.

Proposition 2. Let $AF = (AR, \text{Attacks})$ and $AF' = (AR', \text{Attacks}')$ be two argumentation frameworks for which it holds that $AF \preceq_N AF'$. For all $g^i \circ \sigma^y_x(AF')$, with $x \in \{\text{complete, grounded, preferred, ideal, semi-stable, eager}\}$ and $y \in \{\cap, \cup\}$, the following statement holds true:

$$g^i \circ \sigma^y_x(AF') \subseteq \text{AR} \iff g^i \circ \sigma^y_x(AF') = g^i \circ \sigma^y_x(AF)$$

Proof. We provide a proof by contra-example. Let us introduce the following argumentation frameworks:

- $AF = (AR, \text{Attacks}) = (\{a, b\}, \{(a, b)\})$;
- $AF' = (AR', \text{Attacks}') = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$.

We can see that $AF \preceq_N AF'$. The argumentation frameworks are resolved as follows:

- $g^i \circ \sigma^y_x(AF) = \{a\}$. 

15
• \( g^i \circ \sigma_x^N(AF') = \{\} \).

It follows that \( g^i \circ \sigma_x^N(AF') \subseteq AR \) and \( g^i \circ \sigma_x^N(AF') \neq g^i \circ \sigma_x^N(AF) \). Consequently, it is clear that \( g^i \circ \sigma_x^N(AF') \subseteq AR \implies g^i \circ \sigma_x^N(AF') = g^i \circ \sigma_x^N(AF) \).

\[ \Phi \]

**Proposition 3.** Let \( AF = (AR, Attacks) \) and \( AF' = (AR', Attacks') \) be two argumentation frameworks for which it holds that \( AF \preceq_N AF' \). For all \( g^i \circ \sigma_x^N \), with \( x \in \{\text{stage}, \text{stage2}, CF2\} \) the following holds true:

\[ g^i \circ \sigma_x^N(AF') \subseteq AR \implies g^i \circ \sigma_x^N(AF') = g^i \circ \sigma_x^N(AF) \]

**Proof.** To provide a proof by contra-example, let us again introduce the following argumentation frameworks:

- \( AF = (AR, Attacks) = (\{a, b\}, \{(a, b)\}) \);
- \( AF' = (AR', Attacks') = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\}) \).

We can see that \( AF \preceq_N AF' \). The argumentation frameworks are resolved as follows:

- \( g^i \circ \sigma_x^N(AF) = \{a\} \).
- \( g^i \circ \sigma_x^N(AF') = \{\} \).

It follows that \( g^i \circ \sigma_x^{\text{cap}}(AF') \subseteq AR \) and \( g^i \circ \sigma_x^{\text{cap}}(AF') \neq g^i \circ \sigma_x^{\text{cap}}(AF) \). Consequently, it is clear that \( g^i \circ \sigma_x^{\text{cap}}(AF') \subseteq AR \implies g^i \circ \sigma_x^{\text{cap}}(AF') = g^i \circ \sigma_x^{\text{cap}}(AF) \).

Note that the proof of Proposition 3 is analogous to the proof of Proposition 2 at large. In words, given \( AR, \{a\} \) is preferred over \( \{\} \) and given \( AR', \{\} \) is preferred over \( \{a\} \). Hence, the semantics do not comply with the reference independence principle; adding an element \( c \) to the set of elements \( \{a, b\} \) can affect the preference an agent has over elements in \( 2^{\{a,b\}} \).

Figure 1 depicts the argumentation graphs of the example frameworks used in the proofs of Proposition 2 and Proposition 3.

By using a different example, we can prove that lenient stage semantics, lenient stage2 semantics, and lenient CF2 semantics do not comply with the reference independence principle, either.

**Proposition 4.** Let \( AF \) and \( AF' \) be argumentation frameworks, with \( AF = (AR, Attacks) \), \( AF' = (AR', Attacks') \), and \( AF \preceq_N AF' \). Let \( g^i \circ \sigma_x^j \) be any argumentation-based choice function with \( x \in \{\text{stage}, \text{stage2}, CF2\} \). The following holds true:

\[ g^i \circ \sigma_x^j(AF) \subseteq AR \implies g^i \circ \sigma_x^j(AF') = g^i \circ \sigma_x^j(AF) \]

**Proof.** We provide a proof by contra-example and introduce the following argumentation frameworks:

- \( AF = (AR, Attacks) = (\{a, b, c\}, \{(a, b)\}) \);
- \( AF' = (AR', Attacks') = (\{a, b, c, d, e\}, \{(a, b), (b, d), (d, c), (c, e), (e, a), (d, d), (e, e)\}) \).
Abstract Argumentation and the Rational Man

Figure 3: Reference dependence: $AF$ implies $\{a, c\}$ is preferred over $\{a, b, c\}$, while $AF'$ implies $\{a, b, c\}$ is preferred over $\{a, c\}$.

We can see that $AF \preceq_N AF'$. Any $\sigma_x^\cup, x \in \{\text{stage}, \text{stage2}, CF2\}$ resolves the frameworks as follows:

1. $g^i \circ \sigma_x^\cup(AF) = \{a, b, c\}$.
2. $g^i \circ \sigma_x^\cup(AF') = \{a, c\}$.

From $g^i \circ \sigma_x^\bot(AF') \subseteq AR$ and $g^i \circ \sigma_x^\bot(AF') \neq g^i \circ \sigma_x^\bot(AF)$ it follows that $g^i \circ \sigma_x^\cup(AF') \subseteq AR \implies g^i \circ \sigma_x^\cup(AF') = g^i \circ \sigma_x^\cup(AF)$.

Figure 3 depicts the argumentation frameworks that are used in the proof above. We can see that $g^i \circ \sigma_x^\cup(AF) = \{a, c\}$ implies $\forall \text{Args} \in AR: \{a, c\} \succeq \text{Args}$, and in particular $\{a, c\} \succeq \{a, b, c\}$, which is inconsistent with $g^i \circ \sigma_x^\cup(AF') = \{a, b, c\}$, which implies $\forall \text{Args} \in AR: \{a, b, c\} \succeq \text{Args}$, and in particular $\{a, b, c\} \succeq \{a, c\}$.

7. An Expansion to Ensure Reference Independence

In the previous section we have shown that none of the evaluated semantics comply with the rational man’s argumentation principles–even when considering lenient semantics. Hence, the results presented in Section 6 indicate that in order to guarantee economic rationality, and in particular reference independence, it is relevant to look beyond argumentation semantics. At first glance, it is striking that the example expansions in the proofs of Propositions 2, 3, and 4 add new cycles to the argumentation graphs. Consequently, we examine if an argumentation framework expansion can be defined that can guarantee compliance with the rational man’s argumentation principle by further restricting the relationship of two argumentation frameworks $AF \preceq_N AF'$.

For this, we first introduce definitions of non-cyclic attack sequences and attack cycles in the context of argumentation frameworks.

**Definition 17** (Non-cyclic Attack Sequences in Argumentation Frameworks). A non-cyclic attack sequence $V$ in an argumentation framework $AF = (AR, Attacks)$ is an argument

---

7. Lenient semantics can be considered to be too lax in that their aggregation of extensions through the union of credulous semantics may combine extensions that contain arguments that do not fulfill the semantics’ axioms (they are not conflict-free); this property is defined by Bench-Capon and Dunne as “multiplicity” (Bench-Capon & Dunne, 2007) and is generally considered as problematic.
sequence \((a_1, ..., a_n)\) such that \((a_i, a_{i+1}) \in \text{Attacks}, a_i \in AR, a_i \neq a_j, 1 \leq j \leq i, 1 \leq i \leq n - 1. \mathcal{V}(AF)\) denotes all non-cyclic attack sequences of \(AF\).

Note that the definition of a non-cyclic attack sequence is useful for a proof we provide below.

**Definition 18** (Attack Cycles in Argumentation Frameworks). An attack cycle \(C\) in an argumentation framework \(AF = (AR, \text{Attacks})\) is an argument sequence \((a_1, a_2, ..., a_{n-1}, a_n)\) such that \((a_i, a_{i+1}) \in \text{Attacks}, a_i \in AR, 1 \leq i \leq n - 1, a_1 = a_n, \text{ and } \forall (a_i, a_{i+1}), (a_j, a_{j+1}) \in C, i \neq j : (a_i, a_{i+1}) \neq (a_j, a_{j+1})\). \(\mathcal{C}(AF)\) denotes all attack cycles of \(AF\) and \(AR^\mathcal{C}\) denotes the arguments that occur in an attack cycle \(C\).

Let us show how cycles are resolved by decision-making functions that use admissible set-based semantics. We can proof that in admissible set-based extensions, all arguments of an odd-length cycle that is not attacked from arguments outside the cycle are not returned by the argumentation-based decision-making function.

**Proposition 5.** Let \(AF = (AR, \text{Attacks})\) be an argumentation framework and let \(\sigma\) be an admissible set-based semantics. If \(\exists c \in \mathcal{C}(AF), |c| \mod 2 = 1 \land \exists a \in \sigma^g(\mathcal{AF}), y \in \{\cap, \cup\}\) such that \((a, b) \in \text{Attacks} \land b \in AR^c\), then \(AR^c \cap g \circ \sigma^y = \emptyset\).

**Proof.** We provide a proof by induction, where \(n = |AR|\).

**Base case:** \(AR^c = AR\). If \(AR^c = AR\), then \(\sigma^y(\mathcal{AF}) = \emptyset\), by definition of an admissible set.

**Induction case:** \(AR^c \neq AR\). If \(b \in AR \setminus AR^c \land b \in \sigma^y(\mathcal{AF}) \implies (b, a) \notin \text{Attacks}\). It follows that all cycles in \(c \in \mathcal{C}(AF)\) can be resolved as shown in the base case.

In the case of any even-length cycle that is not attacked from arguments outside the cycle, an argumentation-based decision-making function that takes the intersection of admissible set-based semantics does not return any arguments within the cycle.

**Proposition 6.** Let \(AF = (AR, \text{Attacks})\) be an argumentation framework and let \(\sigma\) be an admissible set-based semantics. If \(\exists c \in \mathcal{C}(AF), |c| \mod 2 = 0 \land \exists a \in \sigma^\cap(\mathcal{AF})\) such that \((a, b) \in \text{Attacks} \land b \in AR^c\), then \(AR^c \cap g \circ \sigma^\cap = \emptyset\).

**Proof.** We provide a proof by induction, where \(n = |AR|\).

**Base case:** \(AR^c = AR\). If \(AR^c = AR\), then \(\sigma^\cap(\mathcal{AF}) = \emptyset\), by definition of an admissible set.

**Induction case:** \(AR^c \neq AR\). If \(b \in AR \setminus AR^c \land b \in \sigma^\cap(\mathcal{AF}) \implies (b, a) \notin \text{Attacks}\). It follows that all cycles \(c \in \mathcal{C}(AF)\) can be resolved as shown in the base case.
Let us observe that in contrast, an argumentation-based decision-making function that takes the union of complete, preferred, stable, semi-stable semantics returns all arguments within the cycle. These results provide further indications that cycles are at the core of the problem of reference dependent argumentation.

Now, we define the concept of a non-cyclic expansion.

**Definition 19 (Non-Cyclic Expansion).** A non-cyclic expansion of two argumentation frameworks $AF = (AR, Attacks)$ and $AF' = (AR', Attacks')$ (denoted by $AF \leq_{NC} AF'$) is an expansion $AF \leq_E AF'$, for which it holds true that $C(AF') = C(AF)$.

This allows us to define the rational man’s expansion.

**Definition 20 (Rational Man’s Expansion).** A rational man’s argumentation expansion of two argumentation frameworks $AF$ and $AF'$ (denoted by $AF \leq_{RM} AF'$) is an expansion $AF \leq E AF'$, for which the following conditions hold:

1. $AF \leq_N AF'$;
2. $AF \leq_{NC} AF'$.

We provide the following proposition–our assumption that the rational man’s expansion guarantees reference independence.

**Proposition 7.** Let $g^i \circ \sigma_y^x$ be an argumentation-based choice function with $x \in \{\text{complete}, \text{preferred}, \text{semi-stable}, \text{stage}, \text{stage2}, \text{CF2}, \text{grounded}, \text{ideal}, \text{eager}\}$, $y \in \{\cap, \cup\}$. Let $AF = (AR, Attacks)$ and $AF' = (AR', Attacks')$ be argumentation frameworks, with $AF \leq_{RM} AF'$. The following statement holds true:

$$g^i \circ \sigma_y^x(AR') \subseteq AR \Longrightarrow g^i \circ \sigma_y^x(AF') = g^i \circ \sigma_y^x(AF)$$

**Proof.** We provide a proof by contradiction. We proof that the following statement does not hold true:

$$g^i \circ \sigma_y^x(AR') \subseteq AR \wedge g^i \circ \sigma_y^x(AF') \neq g^i \circ \sigma_y^x(AF)$$

Let us introduce the following observation:

If $g^i \circ \sigma_y^x(AR') \subseteq AR \wedge g^i \circ \sigma_y^x(AF') \neq g^i \circ \sigma_y^x(AF)$

$$\Longrightarrow g^i \circ \sigma_y^x(AR') \subseteq AR \wedge$$

$$(g^i \circ \sigma_y^x(AF') \not\subseteq g^i \circ \sigma_y^x(AF) \lor$$

$$g^i \circ \sigma_y^x(AF') \not\subseteq g^i \circ \sigma_y^x(AF'))$$

In this observation, there are two cases:

1. If $g^i \circ \sigma_y^x(AF') \not\subseteq g^i \circ \sigma_y^x(AF)$:

$$\Longrightarrow \exists a \in g^i \circ \sigma_y^x(AF') \wedge a \not\in g^i \circ \sigma_y^x(AF) \wedge a \in AR$$

$$\Longrightarrow \exists (b, ..., a) \in V(AF'), b \in g^i \circ \sigma_y^x(AF')$$

$$\Longrightarrow \exists b \in AR' \setminus AR, b \in g^i \circ \sigma_y^x(AF')$$

$$\Longrightarrow g^i \circ \sigma_y^x(AF') \not\subseteq AR.\ This\ is\ a\ contradiction.$$
2. If \( g^i \circ \sigma_x^\cup(AF) \not\subseteq g^i \circ \sigma_x^\cup(AF') \):

\[
\begin{align*}
&\implies \exists a \in g^i \circ \sigma_x^\cup(AF) \land a \not\in g^i \circ \sigma_x^\cup(AF') \land a \in AR \\
&\implies \exists (b, \ldots, a) \in V(AF'), b \in g^i \circ \sigma_x^\cup(AF') \\
&\implies \exists b \in AR' \setminus AR, b \in g^i \circ \sigma_x^\cup(AF') \\
&\implies g^i \circ \sigma_x^\cup(AF') \not\subseteq AR. \text{This is a contradiction.}
\end{align*}
\]

In the context of the proof, it is worth noting that by definition, for stage2 and CF2 semantics, given \( AF \preceq_{NC} AF' \), it holds for any \( SCC_{AF'}, SCC_{AF} \not\subseteq AF \) that \( |SCC_{AF'}| = 1 \), i.e. any strongly connected component that is in \( AF' \) and not in \( AF \) consists of exactly one argument.

Let us note that allowing for expansions that add even cycles to an argumentation framework cannot guarantee reference independence in the case of many skeptical semantics. For example, given \( AF = (\{a,b\}, \{(a,b)\}) \) and its expansion \( AF' = (\{a,b,c,d\}, \{(a,b),(b,c),(c,d),(d,a)\}) \), grounded, ideal, and eager semantics, as well as skeptical complete, preferred, semi-stable, stage, stage2, and CF2 semantics return \( \{\{a\}\} \) for \( AF \) and \( \{\}\) for \( AF' \), as depicted in Figure 4. Also, it is not sufficient that only cycles that include at least one argument \( Arg \in AR \) and at least one argument \( Arg \in AR' \setminus AR \) are not allowed in an expansion \( AF \preceq_{NC} AF' \). This can be shown by introducing the following example. Given \( AF = (\{a\}, \{\}) \) and \( AF' = (\{a,b,c,d\}, \{(b,a),(c,b),(d,c),(b,d)\}) \), grounded, ideal, and eager semantics, as well as skeptical complete, preferred, semi-stable, stage, stage2, and CF2 semantics return \( \{\{A\}\} \) for \( AF \) and \( \{\}\) for \( AF' \), as depicted in Figure 5. Given the presented findings, it is obvious that if expansion and reduction (the removal of arguments) of an argumentation framework are allowed in any argumentation scenario, cycles should be avoided altogether. For example, the argumentation framework \( AF = (\{a,b,c\}, \{(a,b),(b,c),(c,a)\}) \) implies (among others) the preference \( \{\} \succeq \{a\} \). Removing the argument \( c \) from \( AF \) gives us \( AF' = (\{a,b\}, \{(a,b)\}) \), with \( AF' \preceq_{N} AF \). As \( AF' \) implies (among others) \( \{a\} \succeq \{\} \), the preferences implied by \( AF \) are inconsistent with the preferences implied by \( AF' \).
8. Rational Man’s Argumentation, Belief Revision, and Dialogues

To highlight the relevance of the presented research, this section provides examples that illustrate how the newly established principles and expansions can be applied.

Example 6 (Single-agent choice). Let us assume we have a rational buyer agent $A$ that frequently (at different points in time $\{t_0,t_1,t_n\}$) chooses which items to buy from a set of products $S_{t_i}$ using an argumentation-based choice function $g$ as defined by Definition 13. At point $t_0$, $A$ chooses from the set $S_{t_0} = \{p_1,p_2\}$ by constructing and resolving the argumentation framework $AF_{t_0} = (S_{t_0},\{\langle p_1,p_2 \rangle \})$, i.e., $A$ chooses $\{p_1\}$. At point $t_1$, $A$ can choose from $S_{t_1} = \{p_1,p_2,p_3\}$. Now, assuming that $A$’s beliefs about $p_1$ and $p_2$ did not change, $A$, as a rational decision-maker, must construct an argumentation framework $AF_{t_1} = (S_{t_1},Attacks_{t_1})$, with $AF_{t_0} \preceq_{NC} AF_{t_1}$. For example, $Attacks_{t_1}$ must not equal $\{\langle p_1,p_2 \rangle,\langle p_2,p_1 \rangle \}$, as then $AF_{t_1}$ cannot be a normal expansion (and consequently not a non-cyclic expansion). With the attacks $\{\langle p_1,p_2 \rangle,\langle p_2,p_3 \rangle,\langle p_3,p_1 \rangle \}$, $AF_{t_1}$ is a normal expansion, but not a non-cyclic expansion of $AF_{t_0}$ (see Figure 6). In contrast, $(S_{t_1},\{\langle p_1,p_2 \rangle,\langle p_3,p_1 \rangle,\langle p_3,p_2 \rangle \})$ is a non-cyclic expansion and hence permitted. If $A$’s beliefs about $p_1$ and/or $p_2$ changed, any change in attack relations can be permitted.

Example 7 (Argumentation dialogues). In a multi-agent scenario, let us assume we have a decision-maker agent $A_1$ that receives advice from a consultant agent $A_2$. In this context, $A_1$ presents its argumentation framework $AF = (AR,Attacks)$ to $A_2$, who then proposes changes by providing $AF' = (AR',Attacks')$, with $AF'$ being a normal expansion of $AF$ ($AF \preceq_N AF'$). Subsequently, $A_1$ can accept or reject the changes. $A_2$ can propose two types of changes:

Set-expanding changes. $A_2$ only shows $A_1$ that additional options to choose from exist and how these options should be integrated into the argumentation framework.

Belief-revising changes. $A_2$ advises $A_1$ to change its beliefs about the choice options contained in $AF$ and may in addition propose set-expanding changes.

---

8. $g$ can make use of any semantics that fulfill the uniqueness principle.
A2 might want to deceive A1 by proposing changes that A2 labels as set-expanding but that are also belief-revising. For example, A1 presents the following argumentation framework to A2:

\[ AF = (\{a, b, c\}, \{(b, a), (c, a)\}) \]

Then, B proposes the following:

\[ AF' = (\{a, b, c, d, e, f\}, \{(b, a), (c, a), (d, c), (e, d)(f, c)(d, f)\}) \]

If A2 labels this proposal as set-expanding, A2 is deceiving A1 to think that A1 is barely considering new options and not revising the assessment of the existing options. However, given the work presented above, A1 can easily detect that \( AF \not\leq_{RM} AF' \) (\( AF \not\leq_{NC} AF' \). Figure 7 depicts the argumentation graphs of AF and AF'.

Example 8 (Argument Mining). The ability to assess the economic rationality, and in particular reference independence, of argumentation frameworks can potentially be useful in argument mining scenarios, in which argumentation graphs are generated from natural language text (Lippi & Torroni, 2016). Let us introduce a scenario where an argument miner uses machine learning techniques for natural language processing to generate argumentation frameworks from text—for example, from legal documents—and then hands them over to an agent that resolves the argumentation frameworks to inform its decision-making. However,
the argumentation agent is not accepting the frameworks under any condition; instead, it is assessing the frameworks and their relation with each other to determine if conditions of economic rationality are infringed. The argumentation agent then provides the results of these assessments to the argument miner, who can use the information in different ways. It can either re-assess the corresponding text and suggest an alternative, economically rational interpretation, or consider the text as not useful and label it accordingly to increase its ability to focus on more useful texts in the future. Figure 8 depicts the architecture of the proposed system. As an example, let us assume the argument miner creates argumentation frameworks based on an evolving online discussion on whether a policy should be implemented or not (denoted by argument \(p\)). At time \(t_0\) the argument miner detects an argument \(a\) that attacks the policy implementation proposal: \(AF_0 = (\{p, a\}, \{(a, p)\})\). The argumentation agent—using, for example, skeptical complete semantics \(\sigma_{\text{complete}}\)—resolves \(AF_0\) to \(\{a\}\), i.e., it decides the policy should not be implemented. At time \(t_1\), the argument miner detects the additional arguments \(b\) and \(c\), as well as the additional attack relations \((a, b)\), \((b, c)\), and \((c, a)\). The argumentation agent resolves the framework \(AF_1 = (\{p, a, b, c\}, \{(a, p), (a, b), (b, c), (c, a)\})\) as \(\{}\). Now, it is clear that \(AF_0 \preceq_N AF_1\), whereas \(AF_0 \not\succeq_{RM} AF_1\). It is also clear that the preference implied by \(g \circ \sigma_{\text{complete}}(AF_0)\) and \(g \circ \sigma_{\text{complete}}(AF_1)\) are inconsistent, i.e., \(g^i \circ \sigma_{\text{complete}}(AF_0) \implies \{a\} \succeq \{}\) and \(g^i \circ \sigma_{\text{complete}}(AF_1) \implies \{} \succeq \{a\}\). Hence, the argumentation agent can label \(AF_1\) as faulty or not useful and provide this information to the argument miner, who can then attempt to find alternative formal interpretations of the discussion, or move on to a different discussion. Figure 9 depicts the argumentation graphs of \(AF_0\) and \(AF_1\).

### 9. Related Work: Preference-based Argumentation and Rational Man’s Expansions

In our work, we derive implicit preferences from Dung-style argumentation frameworks. Consequently, it makes sense to put our work in the context of argumentation approaches that explicitly define preferences. Amgoud’s and Cayrol’s preference-based argumenta-
tion (Amgoud & Cayrol, 2002) can be considered the most foundational work that advances this research direction. Hence, we relate our work to preference-based argumentation and confirm the intuition that the explicit definition of preferences does not guarantee rationality by formal proof. Let us first introduce a definition of a preference-based argumentation framework.

**Definition 21** (Preference-based Argumentation Framework). (Amgoud & Cayrol, 2002) A preference-based argumentation framework is a triplet \((\mathcal{AR}, \text{Attacks}, \text{Prefs})\), whereby \(\mathcal{AR}\) and \(\text{Attacks}\) are arguments and attack relations, defined according to Definition 4 and \(\text{Prefs}\) define a partial or total ordering over \(\mathcal{AR} \times \mathcal{AR}\).

In a preference-based argumentation framework, acceptability is determined as follows.

**Definition 22** (Preference-based Argumentation Framework). (Amgoud & Cayrol, 2002) Given a preference argumentation framework \(\mathcal{AF}_p = (\mathcal{AR}, \text{Attacks}, \text{Prefs})\), the set of acceptable arguments \(\text{Args}_{\text{acc}} \subseteq \mathcal{AR}\) is determined as follows by the preference-based argumentation function \(\tau_{\text{pref}}:\)

\[
\tau_{\text{pref}}(\mathcal{AF}) = \{a \in \mathcal{AR}| \forall b \in \mathcal{AR} \text{ if } (b,a) \in \text{Attacks then } a \geq b\}
\]

To analyze preference-based argumentation in the context of reference independence, let us first define the concept of a normal expansion of a preference-based argumentation framework.

**Definition 23.** An expansion \(\mathcal{AF}'_p = (\mathcal{AR}', \text{Attacks}', \text{Prefs}')\) of an argumentation framework \(\mathcal{AF} = (\mathcal{AR}, \text{Attacks}, \text{Prefs})\) is normal \((\mathcal{AF}_p \preceq_{NP} \mathcal{AF}'_p)\) iff:

- \(\forall a, b : ((a, b) \in \text{Attacks}' \Rightarrow a \in \mathcal{AR}' \\setminus \mathcal{AR} \lor b \in \mathcal{AR}' \setminus \mathcal{AR})\) and
- \(\forall (a \geq b) \in \text{Prefs} \Rightarrow (a \geq b) \in \text{Prefs}'\) and
- \(\forall a, b : ((a \geq b) \in \text{Prefs}' \setminus \text{Prefs} \Rightarrow a \in \mathcal{AR}' \setminus \mathcal{AR} \lor b \in \mathcal{AR}' \setminus \mathcal{AR})\).

In words, considering the addition of preferences \(\text{Prefs}\) to Dung-style argumentation frameworks, we assume that a normal expansion \(\mathcal{AF}'_p\) of \(\mathcal{AF}_p\) does neither change existing preferences defined in \(\text{Prefs}\) nor add additional preferences between any two arguments that exist in \(\mathcal{AF}_p\). Normal and non-cyclic expansions do not require a definition that is specific to preference-based argumentation.
Abstract Argumentation and the Rational Man

Figure 10: Normal, but cyclic expansion: $AF_p \preceq_{NP} AF'_p$, but $AF_p \not\preceq_{NC} AF'_p$.

Now, it can be easily shown that preference-based argumentation does not guarantee reference independence. For this, we introduce the following proposition.

**Proposition 8.** Let $AF_p$ and $AF'_p$ be argumentation frameworks, with $AF_p = (AR, Attacks, Prefs)$, $AF'_p = (AR', Attacks', Prefs')$, and $AF_p \preceq_{NP} AF'_p$. Let $g^i \circ \tau_{pref}$ be an argumentation-based choice function that uses preference-based argumentation. The following holds true:

$$g^i \circ \tau_{pref}(AF'_p) \subseteq AR \iff g^i \circ \tau_{pref}(AF'_p) = g^i \circ \tau_{pref}(AF_p)$$

**Proof.** The proposition can be proven by contra-example. We introduce the following preference-based argumentation frameworks:

- $AF_p = \{(a, b, c), ((a, b), (a \succeq c, b \succeq c))\}$;
- $AF'_p = \{(a, b, c, d), ((a, b, (b, d), (d, a), (a \succeq c, b \succeq c, d \succeq c))\}$.

We can see that $AF_p \preceq_{NP} AF'_p$. $g^i \circ \tau_{pref}$ resolves the frameworks as follows:

1. $g^i \circ \tau_{pref}(AF_p) = \{a, c\}$.
2. $g^i \circ \tau_{pref}(AF'_p) = \{c\}$.

From $g^i \circ \tau_{pref}(AF'_p) \subseteq AR$ and $g^i \circ \tau_{pref}(AF'_p) \neq g^i \circ \tau_{pref}(AF_p)$ it follows that $g^i \circ \tau_{pref}(AF'_p) \subseteq AR \implies g^i \circ \tau_{pref}(AF'_p) = g^i \circ \tau_{pref}(AF_p)$.  

Figure 10 depicts the argumentation frameworks used in the proof.

Also, it can be proven that even when $AF'_p$ is a normal, non-cyclic expansion of $AF_p$ ($AF_p \preceq_{NP} AF'_p, AF_p \preceq_{NC} AF'_p$), reference dependence is not guaranteed.

**Proposition 9.** Let $AF_p$ and $AF'_p$ be argumentation frameworks, with $AF_p = (AR, Attacks, Prefs)$, $AF'_p = (AR', Attacks', Prefs')$, and $AF_p \preceq_{NP} AF'_p, AF_p \preceq_{NC} AF'_p$. Let $g^i \circ \tau_{pref}$ be any argumentation-based choice function that uses preference-based argumentation. The following holds true:

$$g^i \circ \tau_{pref}(AF'_p) \subseteq AR \implies g^i \circ \tau_{pref}(AF'_p) = g^i \circ \tau_{pref}(AF_p)$$
Proof. Again, the proposition can be proven by contra-example. We introduce the following preference-based argumentation frameworks:

- \( \mathcal{AF}_p = (\{a, b, c\}, \{(a, b)\}, (a \succeq c, b \succeq c)) \);
- \( \mathcal{AF}'_p = (\{a, b, c, d\}, \{(a, b), (c, d), (d, a)\}, (a \succeq c, b \succeq c, d \succeq c)) \).

We can see that \( \mathcal{AF}_p \preceq_{NP} \mathcal{AF}'_p \) and \( \mathcal{AF}_p \preceq_{NC} \mathcal{AF}'_p \). \( g^i \circ \tau_{\text{pref}}(\mathcal{AF}_p) = \{a, c\} \).

\( g^i \circ \tau_{\text{pref}}(\mathcal{AF}'_p) = \{c\} \).

From \( g^i \circ \tau_{\text{pref}}(\mathcal{AF}'_p) \subseteq AR \) and \( g^i \circ \tau_{\text{pref}}(\mathcal{AF}'_p) \neq g^i \circ \tau_{\text{pref}}(\mathcal{AF}_p) \) it follows that \( g^i \circ \tau_{\text{pref}}(\mathcal{AF}'_p) \subseteq AR \not\Rightarrow g^i \circ \tau_{\text{pref}}(\mathcal{AF}'_p) = g^i \circ \tau_{\text{pref}}(\mathcal{AF}_p) \). \( \square \)

The argumentation frameworks used in the proof are depicted by Figure 11. It is clear that preference-based argumentation complies with the principle, given \( \mathcal{AF}_p = (AR, \text{Attacks}, \text{Prefs}) \), \( \mathcal{AF}'_p = (AR', \text{Attacks}', \text{Prefs}') \), and \( \mathcal{AF}_p \preceq_{NP} \mathcal{AF}_p \) and given that \( \text{Prefs} \) and \( \text{Prefs}' \) are strict total orderings. In this case, the preferences over arguments override their attacks relation and render the attack relations meaningless. As preference-based argumentation is a special case of value-based argumentation as introduced by Bench-Capon (Bench-Capon, 2003), it is also clear normal, non-cyclic expansions cannot guarantee rationality for value-based argumentation.

10. Conclusion and Future Work

This paper shows that abstract argumentation semantics typically do not guarantee a rational choice process according to the rational man paradigm and consequently uncovers a gap between abstract argumentation semantics and economically rational choice. Thereby, our research focuses on foundational work at the intersection of abstract argumentation and (bounded) economic rationality. While semantics typically do not comply with rational man’s argumentation principles that this paper establishes, the defined notion of a rational
Abstract Argumentation and the Rational Man

man’s argumentation expansion allows to ensure rationality in the context of argumentation-based choice functions that use Dung-style semantics. Our work provides answers to the question to what extent and under which circumstances the application of abstract argumentation returns “reasonable” results from a practical perspective, in particular from the perspective of economic rationality. Taking into account the rich body of formal works on both argumentation and (boundedly) rational choice, plenty of opportunities to extend our work exist. In particular, we consider the following research directions as promising future work:

- **'Loop-busting’ to ensure economic rationality in temporal argumentation.**
  In this paper, we have shown that Dung-style argumentation approaches are typically economically not rational when considering the normal expansion of argumentation frameworks and that economic rationality—i.e. reference independence—can be achieved by avoiding the addition of cycles in normal expansions. However, it can be assumed that many scenarios require well-defined approaches to handling argumentation cycles in an economically rational (reference independent) manner. To devise such approaches, it should be possible to build on a fundament of works on the resolution of cycles in argumentation graphs (also called ‘loop-busting’) (Baroni, Gabbay, & Giacomin, 2014).

- **Economic rationality and advanced argumentation frameworks.**
  In addition to preference-based and value-based argumentation, a range of other works extends Dung’s notion of an argumentation framework, for example by assigning weights or intervals to attack relations (e.g., probabilistic (Li, Oren, & Norman, 2011) and possibilistic (Nieves & Confalonieri, 2011) argumentation). Given that the rational man’s expansion as established in this paper does not guarantee reference independence in the case of preference-/value-based argumentation, the exploration of the intersection of these approaches and economic rationality can be considered promising future research.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We thank Dov Gabbay for valuable feedback and discussions.

Appendix. Uniqueness in Argumentation Semantics

In addition to showing which argumentation-based choice functions comply with the clear preferences property, we provide an overview of which argumentation semantics can fulfill this property by themselves, without the need of aggregation by $g^U$ or $g^\cap$. We refer to semantics with this behavior as semantics that fulfill the uniqueness principle. By providing examples that return ambiguous results (i.e. more than one extension of a specific semantics for the same argumentation framework), it can be easily proven by example that complete,
Table 1: Overview: compliance with the uniqueness principle.

|                  | Credulous ($\sigma(AF)$) | Skeptical ($\sigma^{\land}(AF)$) | Lenient ($\sigma^{\lor}(AF)$) |
|------------------|--------------------------|----------------------------------|------------------------------|
| Complete         | No                       | Yes                              | Yes                          |
| Grounded         | Yes                      | -                                | -                            |
| Preferred        | No                       | Yes                              | Yes                          |
| Stable           | No                       | No                               | No                           |
| Ideal            | Yes                      | -                                | -                            |
| Semi-stable      | No                       | Yes                              | Yes                          |
| Eager            | Yes                      | -                                | -                            |
| Stage            | No                       | Yes                              | Yes                          |
| CF2              | No                       | Yes                              | Yes                          |
| Stage2           | No                       | Yes                              | Yes                          |

preferred, stable, semi-stable, eager, and stage semantics do not comply with the uniqueness principle.

**Proposition 10.** The following semantics do not comply with the uniqueness principle: complete semantics, preferred semantics, stable semantics\(^9\), semi-stable semantics, and stage semantics.

**Proof.** Let us introduce the argumentation framework $AF_1 = (\{A, B\}, \{(A, B), (B, A)\})$. When applying any of the semantics listed in Proposition 10, $AF_1$ has two extensions: $\{A\}$ and $\{B\}$. \hfill \(\Box\)

Grounded semantics comply with the uniqueness principle. To assert this, one can rely on Dung’s proof that any argumentation framework has exactly one grounded extension (Dung, 1995). For ideal semantics, the proof that every argumentation framework has exactly one (ideal) extension is provided by Caminada and Pigozzi (Caminada & Pigozzi, 2011); the same is proven for eager semantics by Caminada (Caminada, 2007). By definition—because they are defined as the union and intersection, respectively, of all extensions an argumentation framework has, given a specific semantics—all lenient and skeptical variants of semantics that always provide at least one extension are unambiguous and hence fulfill the uniqueness principle; i.e., all skeptical and lenient semantics with the exception of stable semantics comply with the uniqueness principle\(^10\).

---

\(^9\) Indeed, as shown by Dung, some argumentation frameworks do not have any stable extension (Dung, 1995).

\(^10\) Stable semantics are an exception here because they are not universally defined, i.e., there exist argumentation frameworks, for which no stable extension can be defined (Baroni et al., 2018).
References

Amgoud, L., & Cayrol, C. (2002). Inferring from inconsistency in preference-based argumentation frameworks. *Journal of Automated Reasoning, 29*(2), 125–169.

Baroni, P., Caminada, M., & Giacomin, M. (2018). Abstract argumentation frameworks and their semantics. In Baroni, P., Gabbay, D., Massimiliano, G., & van der Torre, L. (Eds.), *Handbook of Formal Argumentation. College Publications*, chap. 4, pp. 159–236.

Baroni, P., Gabbay, D. M., & Giacomin, M. (2014). Introduction to the special issue on Loops in Argumentation. *Journal of Logic and Computation, 26*(4), 1051–1053.

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

Bateman, I., Munro, A., Rhodes, B., Starmer, C., & Sugden, R. (1997). A test of the theory of reference-dependent preferences. *The quarterly journal of economics, 112*(2), 479–505.

Baumann, R., & Woltran, S. (2016). The role of self-attacking arguments in characterizations of equivalence notions. *Journal of Logic and Computation, 26*(4), 1293–1313.

Baumann, R., & Brewka, G. (2010). Expanding argumentation frameworks: Enforcing and monotonicity results. *COMMA, 10*, 75–86.

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

Bench-Capon, T. J. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation, 13*(3), 429–448.

Caminada, M. (2007). Comparing two unique extension semantics for formal argumentation: ideal and eager. In *Proceedings of the 19th Belgian-Dutch conference on artificial intelligence (BNAIC 2007)*, pp. 81–87. Utrecht University Press.

Caminada, M. (2019). Rationality postulates: applying argumentation theory for non-monotonic reasoning.

Caminada, M., & Pigozzi, G. (2011) *Autonomous Agents and Multi-Agent Systems, 22*(1), 64–102.

Caminada, M. W., Carnielli, W. A., & Dunne, P. E. (2012). Semi-stable semantics. *Journal of Logic and Computation, 22*(5), 1207–1254.

Davey, B. A., & Priestly, H. A. (2002). *Introduction to Lattices and Order* (Second edition). Cambridge University Press.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M. S., Lopez-Sanchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman, M., van Steenbergen, M., Tedeschi, S., van der Torre, L., Villata, S., & de Wildt, T. (2018). Ethics by design: Necessity or curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’18*, pp. 60–66, New York, NY, USA. ACM.
Doutre, S., & Mengin, J. (2004). On sceptical versus credulous acceptance for abstract argument systems. In Alferes, J. J., & Leite, J. (Eds.), Logics in Artificial Intelligence, pp. 462–473, Berlin, Heidelberg. Springer Berlin Heidelberg.

Doyle, J. R., O’Connor, D. J., Reynolds, G. M., & Bottomley, P. A. (1999). The robustness of the asymmetrically dominated effect: Buying frames, phantom alternatives, and in-store purchases. Psychology & Marketing, 16(3), 225–243.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. Artificial intelligence, 77(2), 321–357.

Dung, P. M., Mancarella, P., & Toni, F. (2007). Computing ideal sceptical argumentation. Artificial Intelligence, 171(10-15), 642–674.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. American economic review, 93(5), 1449–1475.

Li, H., Oren, N., & Norman, T. J. (2011). Probabilistic argumentation frameworks. In International Workshop on Theorie and Applications of Formal Argumentation, pp. 1–16. Springer.

Lippi, M., & Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. ACM Trans. Internet Technol., 16(2), 10:1–10:25.

Nash, J. F. (1950). Equilibrium points in n-person games. Proceedings of the National Academy of Sciences, 36(1), 48–49.

Nieves, J. C., & Confalonieri, R. (2011). A possibilistic argumentation decision making framework with default reasoning. Fundamenta Informaticae, 113(1), 41–61.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., ..., & Wellman, M. P. (2019). Machine behaviour. Nature, 568, 477–486.

Rahwan, I., & Simari, G. R. (2009). Argumentation in Artificial Intelligence (1st edition). Springer Publishing Company, Incorporated.

Rubinstein, A. (1998). Modeling bounded rationality. MIT press.

Russell, L. (1938). A critical exposition of the philosophy of Leibniz. Philosophy, 13(50), 217–220.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.,

Sen, A. (1991). Utility: Ideas and terminology. Economics and Philosophy, 7(2), 277–283.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. The Quarterly Journal of Economics, 69(1), 99–118.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. The American Economic Review, 80(1), 204–217.

van der Torre, L., & Vesic, S. (2017). The principle-based approach to abstract argumentation semantics. IJCoLog Journal of Logics and Their Applications, 4(8).
Verheij, B. (1996). Two approaches to dialectical argumentation: admissible sets and argumentation stages. *Proc. NAIC*, 96, 357–368.

Von Neumann, J., Morgenstern, O., & Kuhn, H. W. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton university press.