Rationalisation of Profiles of Abstract Argumentation Frameworks: Characterisation and Complexity

Airiau, S.; Bonzon, E.; Endriss, U.; Maudet, N.; Rossit, J.

DOI
10.1613/jair.5436

Publication date
2017

Document Version
Final published version

Published in
Journal of Artificial Intelligence Research

Journal of Artificial Intelligence Research 60 (2017) 149–177 Submitted 01/17; published 09/17

Rationalisation of Profiles of Abstract Argumentation Frameworks: Characterisation and Complexity

Stéphane Airiau STEPHANE.AIRIAU@DAUPHINE.FR
LAMSADE, Université Paris-Dauphine
PSL Research University
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16, France

Elise Bonzon ELISE.BONZON@PARISDESCARTES.FR
LIPADE, Université Paris Descartes
45 rue des Saints Pères, 75006 Paris, France

Ulle Endriss ULLE.ENDRISS@UVA.NL
ILLC, University of Amsterdam
Science Park, 1090 GE Amsterdam, The Netherlands

Nicolas Maudet NICOLAS.MAUDET@LIP6.FR
LIP6, UPMC Université Paris 6, Sorbonne Universités
4 place Jussieu, 75005 Paris, France

Julien Rossit JULIEN.ROSSIT@PARISDESCARTES.FR
LIPADE, Université Paris Descartes
45 rue des Saints Pères, 75006 Paris, France

Abstract

Different agents may have different points of view. Following a popular approach in the artificial intelligence literature, this can be modelled by means of different abstract argumentation frameworks, each consisting of a set of arguments the agent is contemplating and a binary attack-relation between them. A question arising in this context is whether the diversity of views observed in such a profile of argumentation frameworks is consistent with the assumption that every individual argumentation framework is induced by a combination of, first, some basic factual attack-relation between the arguments and, second, the personal preferences of the agent concerned regarding the moral or social values the arguments under scrutiny relate to. We treat this question of *rationalisability* of a profile as an algorithmic problem and identify tractable and intractable cases. In doing so, we distinguish different constraints on admissible rationalisations, e.g., concerning the types of preferences used or the number of distinct values involved. We also distinguish two different semantics for rationalisability, which differ in the assumptions made on how agents treat attacks between arguments they do not report. This research agenda, bringing together ideas from abstract argumentation and social choice, is useful for understanding what types of profiles can reasonably be expected to occur in a multiagent system.

1. Introduction

The model of abstract argumentation introduced by Dung (1995) is at the root of a vast amount of work in artificial intelligence. In a nutshell, this model abstracts away from the internal structure of an argument and simply represents argumentation frameworks as directed

graphs, where the nodes are arguments and the edges are attacks between arguments—in the sense that one argument undercuts or contradicts another argument. Different semantics provide principled approaches to selecting sets of arguments that can be viewed as coherent when advanced together. The simplicity and generality of this framework, as well as its links with nonmonotonic reasoning, have stimulated a number of directions of research, e.g., at the level of the definition of the semantics, of their computation, of the expressivity of such frameworks, or regarding their application in a multiagent system.

In recent years, a number of authors have addressed the problem of aggregating several argumentation frameworks, each associated with the stance taken by a different individual agent, into a single collective argumentation framework that would appropriately represent the views of the group as a whole. Examples include the contributions of Coste-Marquis, Devred, Konieczny, Lagasquie-Schiex, and Marquis (2007), Tohmé, Bodanza, and Simari (2008), Bodanza and Auday (2009), and Dunne, Marquis, and Wooldridge (2012). Aggregating argumentation frameworks is a form of graph aggregation (Endriss & Grandi, 2017): We are given a profile of attack-relations, one for each agent, and are asked to compute a suitable compromise attack-relation. This is an interesting and fruitful line of research, bringing together concerns in abstract argumentation with the methodology of social choice theory,[1] but it raises one important question: For a given profile of argumentation frameworks, is it in fact conceivable that such a profile would manifest itself? That is, how do we explain the differences in perspective of the individual agents for a given profile? Why do they sometimes report different arguments? And why do they sometimes report different attacks even between those arguments they agree on? In this paper, we propose a formal model for studying such questions.

The point that the attack-relation should not be viewed as absolute and objective, but may very well depend on the individual circumstances of the agent considering the arguments in question, has been made before by multiple authors (e.g., Bench-Capon, Doutre, & Dunne, 2007; Amgoud, Dimopoulos, & Moraitis, 2008; Baumann, 2012; Booth, Kaci, & Rienstra, 2013; Grossi & van der Hoek, 2013; Gabbriellini & Torroni, 2013). A widespread explanation for such diversity of views is that agents have different preferences regarding the arguments at hand. For instance, arguments may come from different sources, which agents may trust to varying degrees. Or the arguments may be attached to different moral or social values, which the agents may prioritise differently. This perspective still assumes an underlying ground truth, which however may be interpreted differently, depending on the agents. The same position is also taken by Searle (2001), who puts the case very clearly:

> "Assume universally valid and accepted standards of rationality, assume perfectly rational agents operating with perfect information, and you will find that rational disagreement will still occur; because, for example, the rational agents are likely to have different and inconsistent values and interests, each of which may be rationally acceptable." (page *xv*)

1. The approach sketched here must be clearly distinguished from a second approach combining abstract argumentation and social choice theory found in the literature, which addresses the question of how to aggregate different extensions (or labellings) for a common argumentation framework. This is the approach of, amongst others, both Caminada and Pigozzi (2011) and Rahwan and Tohmé (2010). Bodanza and Auday (2009) compare the two approaches explicitly. We point the reader to the recent survey by Bodanza, Tohmé, and Auday (2017) for a detailed description of all of these works.

In the literature on abstract argumentation, frameworks for modelling this phenomenon have been proposed by several authors, including both Amgoud and Cayrol (2002) and Bench-Capon (2003). Here we adopt a preference-based approach, in the *value-based* variant due to Bench-Capon (2003). In his model, whether argument A ultimately defeats argument B does not only depend on whether A attacks B in an objective sense, but also on how we rank the importance of the moral or social values attached to A and B: If we rank the value associated with B strictly above that associated with A, we may choose to ignore any attacks of A on B. Thus, differences in their preferences can explain why different agents may report different attacks.

Regarding the fact that agents may also report different sets of arguments to begin with, the most natural explanation is simply that the agents are not all *aware* of the same arguments. (We shall mostly stick to this interpretation in this paper). However, depending on the context, it may sometimes also be reasonable to assume that an agent *chooses*, on purpose, not to report certain arguments. For instance, it may be the case that certain values are 'taboo' for some agents, and that they prefer not to refer to them and thus choose to suppress any arguments relating to those values.[2] Or agents may choose to ignore arguments they consider irrelevant, with the aim of minimising communication.

At the technical level, the question we ask in this paper thus is the following: Given a profile of argumentation frameworks $(AF_1, \ldots, AF_n)$, one for each agent, defined over possibly different sets of arguments, can this profile be explained in terms of a single master argumentation framework, an association of arguments with values, and a profile of preference orders over values $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$, one for each agent? Or, as we shall put it: Can the profile of argumentation frameworks observed be *rationalised*? To be able to answer this question in the affirmative, for every agent i, we require AF_i to be exactly the argumentation framework we obtain when the master argumentation framework with its associated values is first restricted to the arguments agent i is aware of and then any attacks that are in conflict with the preference order $\succcurlyeq_i$ are being cancelled. We are also going to consider an alternative notion of rationalisability, where we assume each agent is aware of all arguments but consciously chooses not to report some of them. In this case, rationalisation is possible, if we can obtain AF_i by first cancelling the attacks in conflict with $\succcurlyeq_i$ and then restricting the resulting attack-relation to the arguments agent i has been observed to report. In both cases, we may impose various constraints on admissible rationalisations. For example, we may make certain assumptions regarding the preferences of agents or we may limit the number of values that may be used for rationalisation.

Of course, alternative justifications for the fact that individual argumentation frameworks may differ could be given instead. The preference-based explanation adopted here is not the only option. In particular, agents may interpret arguments differently, especially when their knowledge is incomplete (Black & Hunter, 2012). Also, while we adopt Bench-Capon's value-based approach as the technical foundation on the basis of which to construct our framework and for which to prove our results, there are alternative models of preference-based argumentation, for instance relying on meta-level argumentation (Modgil, 2009). We do not wish to commit to one specific view on the complex question of how to best model preferences in argumentation—see the works of Amgoud and Vesic (2011) and

2. Similar ideas have been explored for the definition of a semantics that attempts to only make use of certain arguments if absolutely necessary (Cayrol, Doutre, Lagasquie-Schiex, & Mengin, 2002).

Modgil and Prakken (2013) for recent contributions to this debate. By confining ourselves to this setting, we favour conceptual simplicity and emphasise our methodological contribution, perhaps at the price of limiting expressivity—more sophisticated argumentation and preference models may score better in this respect and thus be better suited to modelling real-world scenarios. Having said this, there are of course several successful applications of value-based argumentation frameworks, be it in the legal domain (Grabmair & Ashley, 2011), the modelling of political debates (Cartwright & Atkinson, 2009), or ecological policy making (Tremblay & Abi-Zeid, 2016). We refer to Atkinson and Bench-Capon (2016) for a recent overview of several additional applications. This combination of conceptual simplicity and relevance to applications makes this setting a perfect candidate to commence the study of the rationalisability of the argumentative stances of a group of agents.

Still, we believe that our general point is relevant beyond such specific modelling choices, and we see our contribution to be first and foremost as a methodological one. The same type of investigation could be undertaken for other models as well.[3] In a sense, this multiplicity of models is precisely what makes our contribution useful: by providing a collection of results that allow checking whether a profile can be rationalised on such grounds, we provide evidence for guiding the modelling process. The good news is that in many—albeit not all—of the cases analysed in this paper, verification of rationalisability can be performed efficiently, even when the assignment of values to arguments is not known beforehand.

The remainder of the paper is organised as follows. Section 2 presents the relevant background regarding value-based argumentation. Section 3 formally introduces the problem of rationalising a given profile of argumentation frameworks provided by a group of agents, and it presents the different types of constraints on solutions we will consider. Section 4 analyses the single-agent case in detail, while Section 5 investigates the multiagent case. Section 6 presents the alternative approach to defining rationalisability, where agents are assumed to choose not to report certain arguments rather than simply not being aware of them. Finally, Section 7 discusses a number of application scenarios and Section 8 concludes with a review of open questions and possible directions for future work.

2. Notation and Terminology

Following Dung (1995), we define an *argumentation framework* (AF) as a binary attack-relation declared over a set of arguments. In this paper, we are going to restrict ourselves to scenarios for which the set of available arguments is finite.

Definition 1 (AF)**.** *An argumentation framework is a pair $AF = \langle Arg, \rightharpoonup \rangle$, where Arg is a finite set of arguments and $\rightharpoonup$, the attack-relation, is an irreflexive binary relation on Arg.*

If $A \rightharpoonup B$ holds for two arguments $A, B \in Arg$, then we say that A *attacks* B.

Example 1. *Pollution is becoming a major health problem in big cities. City councils are facing the question of possibly banning polluting vehicles, and specifically diesel cars, from the city centres. A city council might entertain the following arguments:*

3. While some preference-based approaches are special cases of the one used here—e.g., in the work of Amgoud and Cayrol (2002) each argument is mapped to a different value—others would require extensions, e.g., allowing several values per argument, as in the work of Kaci and van der Torre (2008).

(A) *Diesel cars should be banned from the inner city center in order to decrease pollution.*

(B) *Artisans, who deserve special protection by the city council, cannot change their vehicles, as that would be too expensive for them.*

(C) *The city can offer financial assistance to artisans.*

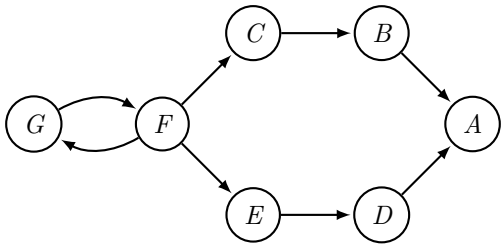
(D) *There are only very few alternatives to using diesel cars. Specifically, the autonomy of electric cars is poor, as there are not enough charging stations around.*

(E) *The city can set up more charging stations.*

(F) *In times of financial crisis, the city should not commit to spending additional money.*

(G) *Health and climate change issues are important, so the city has to spend what is needed to tackle pollution.*

The following graph shows the AF generated by these arguments, together with a natural attack-relation $\rightharpoonup$ between them:



Observe that for this AF it is ambiguous whether or not we should accept argument A and ban diesel cars: Accepting either $\{A, C, E, G\}$ or $\{B, D, F\}$ is intuitively admissible.

Next, we introduce *preferences*. Recall that a *preorder* is a binary relation that is reflexive and transitive, and a *weak order* in addition is also complete (Roberts, 1979). We use preorders and weak orders to model preferences. Using a preorder means allowing for strict preferences, indifferences, and incomparabilities, while using a weak order excludes the possibility of two items being incomparable. We will use the terms 'preference order' and 'preorder' synonymously, i.e., a 'complete preference order' refers to a weak order. The strict part of a preference order $\succcurlyeq$ is denoted as $\succ$ and its indifference part as $\sim$. Thus, we write $x \succ y$ if $x \succcurlyeq y$ but not $y \succcurlyeq x$, and we write $x \sim y$ if both $x \succcurlyeq y$ and $y \succcurlyeq x$.

Following Bench-Capon (2003), we define an *audience-specific value-based argumentation framework* (AVAF) as an AF equipped with a function associating each argument with the social or moral value it advances, combined with a preference order declared over those values. While the mapping from arguments to values is fixed and the same for everyone, the preferences over values are those of a particular agent (the "audience").

Definition 2 (AVAF)**.** *An audience-specific value-based argumentation framework is defined as a 5-tuple $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$, where $\langle Arg, \rightharpoonup \rangle$ is an argumentation framework, Val is a finite set of values, $val : Arg \to Val$ is a mapping from arguments to values, and $\succcurlyeq$ is the audience's preference order on Val.*

We call $\langle Val, val \rangle$ the AVAF's *value-labelling*. Let $=_{val}$ be the equivalence relation on arguments induced by val: $A =_{val} B$ if and only if $val(A) = val(B)$.

Now suppose an agent is presented with an AF and a value-labelling. In Bench-Capon's model, this agent will uphold a proposed attack $A \rightharpoonup B$ and therefore accept that A *defeats* B, unless she strictly prefers the value associated with the attackee B to the value associated with the attacker A (Bench-Capon, 2003).

Definition 3 (Defeated Arguments)**.** *Given an AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$, we say that argument $A \in Arg$ defeats argument $B \in Arg$, denoted $A \rightleftharpoons B$, if and only if we have $A \rightharpoonup B$ but it is not the case that $val(B) \succ val(A)$.*

We call $\rightleftharpoons$ the defeat-relation *induced* by the AVAF. We stress that saying 'it is not the case that $val(B) \succ val(A)$' is the same as saying '$val(A) \succcurlyeq val(B)$ is the case' only when the preference order $\succcurlyeq$ is complete.

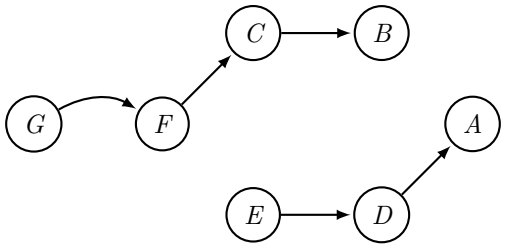
Note that for any given AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$ the induced defeat-relation $\rightleftharpoons$ is, just like an attack-relation $\rightharpoonup$, an irreflexive binary relation on Arg. Thus, we can (and will) think of $\langle Arg, \rightleftharpoons \rangle$ as just another AF.

Example 1 (continued)**.** *Recall our earlier example about the arguments pondered by our city council. We can associate the arguments presented in this example with four types of values. Arguments A and G concern environmental responsibility (value* `env`*), B and C are about social fairness (value* `soc`*), F promotes economic viability (value* `econ`*), and D and E pertain to infrastructure efficiency (value* `infra`*). We thus have that $Val = \{\texttt{env}, \texttt{soc}, \texttt{econ}, \texttt{infra}\}$, as well as that $val(A) = val(G) = \texttt{env}$, $val(B) = val(C) = \texttt{soc}$, $val(F) = \texttt{econ}$, and $val(D) = val(E) = \texttt{infra}$.*

Let us now assume that a particular councillor wants to promote the values of environmental responsibility and infrastructure efficiency over the other two values. So her preferences might be given by the following weak order:

$$\texttt{env} \sim \texttt{infra} \succ \texttt{soc} \sim \texttt{econ}$$

This induces a defeat-relation $\rightleftharpoons$ for our councillor that corresponds to the following graph:



For instance, the attack from B to A got removed, because $val(A) = \texttt{env} \succ \texttt{soc} = val(B)$. Overall, three attacks got removed. For the new AF it is unambiguously clear that A should be accepted (the only argument attacking A is itself attacked by an argument without any remaining attackers), and thus that diesel cars should be banned from the city centre.

In the sequel, we are going to use standard set-theoretical operations (e.g., $\cap$, $\subseteq$) on binary relations (understood as sets of pairs). Furthermore, $R^{-1} = \{(x, y) \mid yRx\}$ is the inverse of a binary relation, R^+ is its transitive closure, and R^* is its reflexive-transitive closure. $R \circ R'$ is the composition of R and R'. We also define $R^+_{val} := (R \cup =_{val})^* \circ R \circ (R \cup =_{val})^*$, which is like the usual transitive closure, except that we can move to arguments with the same value, even if not connected by R. Finally, for any binary relation R defined on some set S, we use $R{\restriction_S} = R \cap (S \times S)$ to denote the restriction of R to S.

3. The Rationalisability Problem

Let $\mathcal{N} = \{1, \ldots, n\}$ be a finite set of *agents* (or *audiences*). Suppose each of these agents supplies us with an AF, not necessarily over the same set of arguments.[4] We call this a *profile* of AF's. As we think of each AF in such a profile as the result of having imposed the corresponding agent's preferences on some underlying master AF, we write individual AF's as $\langle Arg_i, \rightleftharpoons_i \rangle$ (rather than as $\langle Arg_i, \rightharpoonup_i \rangle$). Here, Arg_i is the set of arguments agent i is *aware* of and $\rightleftharpoons_i$ is the defeat-relation on Arg_i adopted by i. A profile of such AF's is denoted as $\boldsymbol{AF} = (\langle Arg_1, \rightleftharpoons_1 \rangle, \ldots, \langle Arg_n, \rightleftharpoons_n \rangle)$. Let $Arg := Arg_1 \cup \cdots \cup Arg_n$ denote the set of all arguments at least one agent is aware of.

Now we may ask whether the profile we observe can be *rationalised* (i.e., whether it can be explained)—in terms of a common master AF and a common value-labelling, together with a profile of preference orders, one for each agent. This question gives rise to the *rationalisability problem* defined next.[5] In fact, we define an entire family of rationalisability problems, parameterised by a set of *constraints* imposed on the solutions admitted. We will soon see several concrete examples for such constraints.

Definition 4 (Rationalisability)**.** *A profile of AF's, $\boldsymbol{AF} = (\langle Arg_1, \rightleftharpoons_1 \rangle, \ldots, \langle Arg_n, \rightleftharpoons_n \rangle)$, is called rationalisable for a given set of constraints, if there exist an attack-relation $\rightharpoonup$ on $Arg = Arg_1 \cup \cdots \cup Arg_n$, a set of values Val with a mapping $val : Arg \to Val$, and a profile $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$ of preference orders on Val, all meeting said constraints, such that, for all agents $i \in \mathcal{N}$ and all arguments $A, B \in Arg_i$, it is the case that $A \rightleftharpoons_i B$ if and only if $A \rightharpoonup B$ but not $val(B) \succ_i val(A)$.*

We refer to $\langle Arg, \rightharpoonup \rangle$ as the *master AF*, and consequently to $\rightharpoonup$ as the *master attack-relation*. Some comments on how to interpret Definition 4 are in order. Given the presumed existence of $\langle Arg, \rightharpoonup \rangle$, $val : Arg \to Val$, and $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$, we think of the observed profile $\boldsymbol{AF} = (\langle Arg_1, \rightleftharpoons_1 \rangle, \ldots, \langle Arg_n, \rightleftharpoons_n \rangle)$ as having come about as the result of the following process. First, each agent $i \in \mathcal{N}$ becomes aware of some subset $Arg_i \subseteq Arg$ of the full set of arguments, and thus of the AF $\langle Arg_i, (\rightharpoonup) \restriction_{Arg_i} \rangle$, i.e., of the restriction of the master attack-relation to the set of arguments she is aware of. Then, in a second step, agent i removes any attacks in this AF that are at odds with her preferences, i.e., we get $A \rightleftharpoons_i B$ for $A, B \in Arg_i$ if and only if $A \; (\rightharpoonup) \restriction_{Arg_i} B$ but not $val(B) \succ val(A)$. Thus, we have made a specific modelling choice when defining rationalisability: We assume that agents *first* choose (possibly unconsciously) the subset of arguments to report, and only *then* reduce the attack-relation defined on that subset according to their individual preferences. Another option would have been to assume that the agents first reduce the master attack-relation according to their own preferences, and then choose a subset of arguments to report (e.g., those that they deem most relevant or significant). We consider the first interpretation more natural

4. A common assumption in the literature on the aggregation of AF's is that every individual agent reports an AF over the exact *same set of arguments* (Bodanza & Auday, 2009; Dunne et al., 2012; Tohmé et al., 2008). Here, we instead follow Coste-Marquis et al. (2007), who have argued that allowing for differences in the individual sets of arguments is more realistic. Note that the case of a single shared set of arguments is covered by our model as a special case.

5. A different problem of rationalisability has recently been proposed by Dunne, Dvořák, Linsbichler, and Woltran (2014): Suppose you are given a set of subsets of arguments. Can these subsets possibly correspond, for a given semantics, to the different extensions of *some* single AF?

and shall adopt it for much of this paper. Nevertheless, in Section 6 we are also briefly going to investigate the alternative semantics sketched here.

Definition 4 can be instantiated for different types of *constraints*. In this paper, we are going to consider the following constraints (but others may be of interest as well):

- the master attack-relation $\rightharpoonup$ may be fixed,
- the value-labelling $\langle Val, val \rangle$ may be fixed,
- the number of values $|Val|$ may be bounded from above by some constant k,
- the preference orders $\succcurlyeq_i$ may be required to be complete.

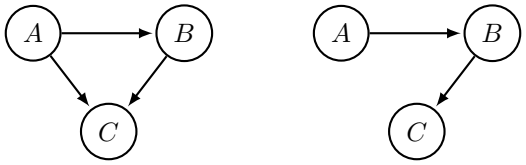
In addition, we will consider one *restriction* of the problem, namely the case where all individual argument sets coincide (i.e., where $Arg_i = Arg_j$ for all $i, j \in \mathcal{N}$). We are also going to treat the single-agent case (with $n = 1$) in some detail.

With these definitions in place, we may now ask: For a given set of constraints, can we characterise the class of all profiles of AF's that can be rationalised? And can we check efficiently whether a given profile is rationalisable?

4. The Single-Agent Case

We first consider the single-agent case of the rationalisability problem. This is not only useful for gaining an understanding of the multiagent case, but is also interesting in its own right. For example, it may be the case that there is some 'ground truth' available and we know what the correct attack-relation is (e.g., due to the logical structure of the arguments), but that a specific agent is still reporting a different AF. Can this subjective AF be explained in terms of the value-based model? That is, is this framework compatible with what we know to be the ground truth?

Example 2. *Consider a scenario with three arguments, $Arg = \{A, B, C\}$, with a fixed master attack-relation $\rightharpoonup$ such that $A \rightharpoonup B$, $B \rightharpoonup C$, and $A \rightharpoonup C$. Suppose we observe a single agent who only declares $A \rightleftharpoons B$ and $B \rightleftharpoons C$. Below, the master attack-relation is shown on the left and the observed individual attack-relation is shown on the right:*



Can we rationalise this omission of the attack of A on C? Clearly, rationalisation requires A and C to be labelled with distinct values, say v_A and v_C, and our agent must prefer v_C to v_A for $A \rightharpoonup C$ to get cancelled. Are two values enough? The answer is no: If we reuse, say, value v_A to also label argument B, then $B \rightharpoonup C$ would get cancelled as well. Similarly, if we reuse v_C for B, then $A \rightharpoonup B$ would get cancelled. Thus, we need a third value v_B. Now there is a rationalisation, with the agent's preference order ranking v_C above v_A, and v_B being incomparable to the other two values. Observe that, even with three values, rationalisation is impossible if we require the preference order to be complete, i.e., if we require it to not leave any two values incomparable.

In the single-agent case, we are given an AF $\langle Arg, \rightrightarrows \rangle$. A solution consists of an AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$, over the same set of arguments Arg, that induces $\rightrightarrows$. In this section, we are going to consider this problem for several types of constraints on solutions. Our aim is to provide polynomial-time algorithms for computing solutions and, where possible, to provide exact characterisations of those solutions. We begin with the simplest of all scenarios, where there are no constraints imposed on permissible rationalisations, and observe that the problem of rationalisability is trivial in this case:

Fact 1 (No Constraints)**.** *In the absence of constraints, every single AF is rationalisable.*

Proof. Given the AF $\langle Arg, \rightrightarrows \rangle$ to be rationalised, let $(\rightharpoonup) := (\rightrightarrows)$, choose the value-labelling $\langle Val, val \rangle$ arbitrarily, and let $(\succcurlyeq) := Val \times Val$, i.e., our agent is indifferent between any two values. Then it is easy to check that $\rightrightarrows$ is induced by the AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$. □

Our proof shows that the same result also applies to rationalisation under any set of constraints referring only to *Val* and *val*. It also continues to apply if we require the preference order to be complete. The main insight here is that any natural instance of the single-agent problem that is nontrivial will involve a constraint on the master attack-relation. Therefore, for the remainder of this section, we only consider rationalisability problems with a given fixed master attack-relation.

Proposition 2 (Fixed Attack-Relation)**.** *A single AF $\langle Arg, \rightrightarrows \rangle$ is rationalisable by an AVAF with a given fixed master attack-relation $\rightharpoonup$ if and only if all of the following three conditions are satisfied:*

(i) $(\rightrightarrows) \subseteq (\rightharpoonup)$*;*
(ii) $(\rightharpoonup \backslash \rightrightarrows)$ *is acyclic;*
(iii) $(\rightrightarrows) \cap (\rightharpoonup \backslash \rightrightarrows)^+ = \varnothing$*.*

Proof. In this setting, there are no constraints on $\langle Val, val \rangle$. The first important insight then is that having more values available means more flexibility: we can rationalise if and only if we can rationalise by labelling every argument with a distinct value. Thus, we may think of the arguments *themselves* as representing values: w.l.o.g., assume that $Val = Arg$ and that *val* is the identity function. Hence, we can think of $\succcurlyeq$ as operating directly on arguments and need not consider values any longer.

Condition (i) is required, as our agent can never add (but only remove) edges. Let $R := (\rightharpoonup \backslash \rightrightarrows)$ denote the set of edges to be removed. We must have $R^{-1} \subseteq (\succ)$ to ensure that the agent's preference order does indeed remove all of these edges. The second important insight now is that it is never beneficial to add more pairs to the preference order than we are absolutely forced to. That is, we should choose $\succ$ as small as possible, namely as the transitive closure of R^{-1}. We then still need to check two things. First, we need to check that $(R^{-1})^+$ is the strict part of some preorder, i.e., that it is transitive and irreflexive. This is equivalent to condition (ii), i.e., to R being acyclic. Second, we need to check that we are not removing any edges that should in fact stay, i.e., we need to make sure that $(\rightrightarrows) \cap R^+ = \varnothing$, which is condition (iii). □

All three conditions can be checked in polynomial time, so we obtain a tractability result:

Corollary 3 (Fixed Attack-Relation)**.** *Whether a single AF is rationalisable by an AVAF with a given fixed master attack-relation can be decided in polynomial time.*

Note that our proof of Proposition 2 shows that requiring the preference order to be strict (i.e., not allowing any indifferences) does not affect rationalisability. On the other hand, our proof does not apply in case the preference order is required to be complete (this case will instead be covered by Proposition 6 below).

As discussed, a crucial ingredient of Proposition 2 and its proof was the fact that there were no constraints on the value-labelling. We now investigate what happens when we add such constraints, and first consider the most extreme case where the full value-labelling is fixed from the outset. This is a natural scenario to consider in those cases in which we are willing to assume that the question of which value a given argument relates to is a matter that can be settled in an objective manner.

Proposition 4 (Fixed Value-Labelling)**.** *A single AF $\langle Arg, \rightleftharpoons \rangle$ is rationalisable by an AVAF with a given fixed master attack-relation $\rightharpoonup$ and a given fixed value-labelling $\langle Val, val \rangle$ if and only if all of the following three conditions are satisfied:*

(i) $(\rightleftharpoons) \subseteq (\rightharpoonup)$*;*
(ii) the relation $\bigcup_{A(\rightharpoonup \setminus \rightleftharpoons)B}\{(val(A), val(B))\}$ *is acyclic;*
(iii) $(\rightleftharpoons) \cap (\rightharpoonup \setminus \rightleftharpoons)^{+}_{val} = \varnothing$*.*

Proof. As for Proposition 2, condition (i) reflects that our agent cannot add new edges. The crucial difference to the scenario of Proposition 2 is that now we cannot remove edges between arguments that are labelled with the same value. Let $R := (\rightharpoonup \setminus \rightleftharpoons)$ be the set of edges we need to remove. At the level of the values, this induces the relation $\bigcup_{(A,B)\in R}\{(val(A), val(B))\}$ mentioned in condition (ii). As before, the best we can do is to choose as small a preference order as possible, so we should use the transitive closure of the inverse of that relation on values. Condition (ii) then amounts to checking that this is indeed a well-formed preference order. Note that acyclicity implies irreflexivity, so we are correctly checking that we are not trying to remove an edge between two arguments labelled with the same value. Finally, we need to check that we are not removing any edges that should stay. This is taken care of by condition (iii). To see this, note that R^{+}_{val} is the set of edges getting removed. □

Also this characterisation immediately provides us with a polynomial-time algorithm. Thus, we obtain the following result:

Corollary 5 (Fixed Value-Labelling)**.** *Whether a single AF is rationalisable by an AVAF with a given fixed master attack-relation and a given fixed value-labelling can be decided in polynomial time.*

The final single-agent scenario we want to consider here is one where we are not given the full value-labelling but merely an upper bound on the number of values that may be used for rationalisation.[6] This scenario comes about when there is no unique objective mapping

6. Thus, this scenario requires solving the decision problem corresponding to the optimisation problem of computing the minimal number of values needed for rationalisation.

from arguments to values and we are looking for a "simple" explanation for an observed defeat-relation only involving a limited number of different values. (For instance, when values correspond to different sources providing information, their number may be known.) From an algorithmic point of view, this is the most demanding problem considered so far. Still, at least for the case of complete preferences, also for this problem we are able to establish the existence of a polynomial-time algorithm, as the following result shows.

Proposition 6 (Bound on Values)**.** *Whether a single AF is rationalisable by an AVAF with a given fixed master attack-relation, a given upper bound on the number of values, and a complete preference order can be decided in polynomial time.*

Proof. We are going to show how to translate our problem into an integer program with at most two variables per inequality. Deciding feasibility of such programs is known to be polynomial (Hochbaum & Naor, 1994).

Let $\langle Arg, \rightleftharpoons \rangle$ be the AF, $\rightharpoonup$ the master attack-relation, and k (with $k \leqslant |Arg|$) the upper bound on the number of values. Observe that, if rationalisation is possible with fewer than k values, then it certainly is possible with exactly k values. As the rationalising preference order is required to be complete, w.l.o.g., we may assume that $Val = \{1, \ldots, k\}$ and that $\succcurlyeq$ is the usual relation $\geqslant$ defined over the natural numbers. Clearly, if $(\rightleftharpoons) \not\subseteq (\rightharpoonup)$, then rationalisation is impossible. So, from now on, assume that $(\rightleftharpoons) \subseteq (\rightharpoonup)$.

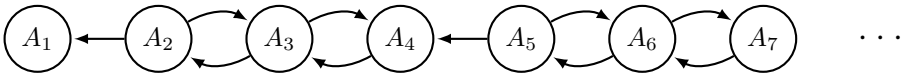
For every argument $A \in Arg$, introduce an integer variable x_A. We use inequalities of the form $1 \leqslant x_A$ and $x_A \leqslant k$ to ensure that each such variable must take a value from Val. Thus, these variables encode val. We have to be able to model two types of constraints. First, if $A \rightharpoonup B$ but not $A \rightleftharpoons B$, then we must ensure that the value of B is strictly preferred to the value of A: $x_A + 1 \leqslant x_B$. Second, if $A \rightleftharpoons B$ (and thus, by our assumption, also $A \rightharpoonup B$), then we must ensure that the value of B is *not* strictly preferred to the value of A: because of completeness, this can be written as $x_B \leqslant x_A$. The integer program thus constructed is feasible if and only if rationalisation is possible. □

Let us reiterate that our proof makes use of the condition that the rationalising preference order should be complete. Without it, we would not be able to map requirements of the form $val(B) \not\succ val(A)$ into linear constraints. Assuming completeness of the preference order (i.e., excluding the possibility of an agent not being able to compare the importance of two given values) is sometimes reasonable, but certainly not always. Whether single-agent rationalisability for a bounded number of values remains polynomial for possibly incomplete preferences is a nontrivial open question of some interest.

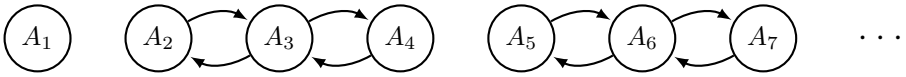
Recall that Example 2 has demonstrated that there indeed are single-agent scenarios where rationalisation is possible with incomplete preferences but impossible with complete preferences. We conclude this section with one further observation on the impact the choice of preference order can have on rationalisability. We show that, even when rationalisation is possible with a complete preference order, imposing that requirement may radically increase the number of values we need to use.

Proposition 7 (Value Ratio)**.** *The ratio between the number of values required to rationalise a given AF by an AVAF with a given fixed master attack-relation and a complete preference order and the number of values required in case the requirement to use a complete preference order is dropped, in the worst case, cannot be bounded from above by a constant.*

Proof. We are going to exhibit a generic example where rationalisation with an incomplete preference order is possible with just three values, while rationalisation with a complete preference order requires $\Omega(|Arg|)$ values. As we can increase the number of arguments arbitrarily, this proves the claim. So, suppose we are given a master attack-relation that consists of a repetition of the same simple gadget of three arguments each (besides A_1):



Now suppose we observe an agent with the following individual defeat-relation, where the first edge in each of the gadgets is missing:



If we permit incomplete preferences, then rationalisation is possible with just three values, $\{v_1, v_2, v_3\}$: To achieve rationalisation, we label the arguments with values in such a way that $val(A_k) = v_{k'}$ whenever $k \equiv k' \bmod 3$, and for the preference order we set $v_1 \succ v_2$ and declare v_3 incomparable to both v_1 and v_2.

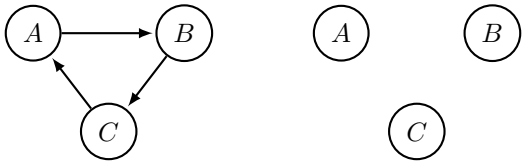
Now suppose we require complete preferences. To preserve the mutual attacks in the middle of each gadget, we must have $val(A_{k+2}) \sim val(A_{k+3}) \sim val(A_{k+4})$ for all $k \geqslant 0$. At the same time, to ensure that the attacks that need to get removed actually do get removed, we must have $val(A_{k+1}) \succ val(A_{k+2})$ for all $k \geqslant 0$. But this is only possible, if we label each gadget with a new value. Therefore, the number of values required for rationalisation is linear in the number of arguments. □

Thus, in principle, the number of values required for rationalisation can grow arbitrarily when we exclude the possibility of an agent declaring two values preferentially incomparable. Of course, the particular AF's used in the proof of Proposition 7 are highly contrived and we should not necessarily expect this growth in the number of values required to be quite that significant in practice. Indeed, a recent experimental study carried out by Greige (2016) shows that, when profiles to be rationalised are generated at random, then exhibiting a scenario where rationalisation is possible with complete preferences but requires strictly more values than rationalisation with arbitrary preferences is extremely rare.

5. The Multiagent Case

We now turn to the multiagent case. In presenting our results for each type of constraint considered, we are specifically going to focus on the extent to which the (positive) results obtained for the single-agent case carry over to this more general scenario. To get started, recall that we have seen that, in the absence of constraints, *every* single AF can be rationalised (Fact 1). The following example shows that this result does not generalise to profiles with (at least) two AF's.

Example 3. *Consider a profile of two AF's over a common set of three arguments. Suppose* $A \rightleftharpoons_1 B$, $B \rightleftharpoons_1 C$, *and* $C \rightleftharpoons_1 A$, *while* $(\rightleftharpoons_2) = \varnothing$, *as shown below:*



Any value-labelled AF and preference profile that could possibly rationalise this profile would have to have an attack-relation $\rightharpoonup$ that includes, at least, the attacks $A \rightharpoonup B$, $B \rightharpoonup C$, and $C \rightharpoonup A$, as otherwise these edges could not have occurred in the first AF. But this means that the second preference order, so as to be able to cancel these attacks, must at least include the comparisons $val(B) \succ_2 val(A)$, $val(C) \succ_2 val(B)$, and $val(A) \succ_2 val(C)$. But then $\succcurlyeq_2$ is not acyclic, thereby contradicting our assumptions on well-defined preference orders. Thus, this profile cannot be rationalised, even in the absence of any kind of constraint.

We are first going to investigate the question of when we can decompose a given multiagent rationalisability problem into a set of n single-agent rationalisability problems that can be solved independently of each other (but that still require us to provide a global solution involving a common master AF and a common value-labelling). Example 3 shows that this kind of decomposition is *not* possible when we do not impose any constraints during rationalisation (i.e., for the scenario covered by Fact 1). On the other hand, for the scenarios of Propositions 2 and 4, it is easy to see that decomposition is possible:

- If the only constraint is that the master attack-relation is fixed, then every agent's rationalisability problem can be solved independently.
- If the only constraints are that the master attack-relation and the value-labelling are fixed, then every agent's rationalisability problem can also be solved independently.

But what if the master attack-relation is not given? Consider the generic profile $\boldsymbol{AF} = (\langle Arg_1, \succcurlyeq_1 \rangle, \ldots, \langle Arg_n, \succcurlyeq_n \rangle)$. Any rationalisation of $\boldsymbol{AF}$ must involve a master attack-relation $\rightharpoonup$ with $(\rightharpoonup) \supseteq (\succcurlyeq_1) \cup \cdots \cup (\succcurlyeq_n)$, because no agent can create an edge not already included in $\rightharpoonup$. Any additional edges in $\rightharpoonup$ will make rationalisation only harder, if they make a difference at all. Thus, rationalisation is possible at all if and only if rationalisation is possible with the fixed master attack-relation $(\rightharpoonup) := (\succcurlyeq_1) \cup \cdots \cup (\succcurlyeq_n)$. Given these insights, together with Corollaries 3 and 5, we obtain the following result:

Proposition 8 (Decomposable Cases)**.** *Whether a profile of AF's is rationalisable can be decided in polynomial time by solving the problem independently for each agent, in at least the following cases:*

(*a*) *No constraints are given.*
(*b*) *The master attack-relation and/or the value-labelling is fixed.*

Thus, of all the constraints we have considered here, only the one specifying an upper bound on the number of values actually leads to a 'genuine' multiagent rationalisation problem. Let us now consider this problem in some detail.

For the remainder of this section, we are always going to assume that a fixed master attack-relation $\rightharpoonup$ is part of the constraints considered. By our reasoning above, any

tractability result obtained under this assumption immediately extends to the case where no master attack-relation is specified.

Our first result on multiagent rationalisation with a bound on the number of values to be used is negative: in the most general case this problem is intractable.

Proposition 9 (Bound on Values: General Case)**.** *Deciding whether a profile of AF's is rationalisable by an AVAF with a given fixed master attack-relation and a given upper bound (of at least 3) on the number of values is an NP-complete problem.*

Proof. NP-membership is immediate. To prove NP-hardness we provide a reduction from Graph Colouring, which is known to be NP-hard (Karp, 1972). Recall that in Graph Colouring we are given an undirected graph $G = \langle V, E\rangle$ and ask whether it is possible to colour the vertices in V using at most $k \geqslant 3$ colours such that no two vertices with the same colour are linked by an edge in E.

So take any instance of Graph Colouring with graph $G = \langle V, E\rangle$ and bound k. Let $m := |V|$. We build an instance of our rationalisation problem for m arguments, $n := \binom{m}{2}$ agents, and a bound of k on the number of values as follows. First, let $Arg := V$ be the full set of arguments, and let the master attack-relation $\rightharpoonup$ be an arbitrary orientation of G. Second, for every pair $A \neq B \in Arg$ we create exactly one agent i, with $Arg_i = \{A, B\}$ and an empty defeat-relation $(\rightrightarrows_i) = \varnothing$. (That is, there indeed are $\binom{m}{2}$ agents.) Now consider any edge (A, B) in G. As either $A \rightharpoonup B$ or $B \rightharpoonup A$, but neither $A \rightrightarrows_i B$ nor $B \rightrightarrows_i A$, the corresponding agent i must strictly rank $val(A)$ and $val(B)$, i.e., they must be different. As this is so for all edges in G and all agents, any two arguments linked in G must get labelled with distinct values. Hence, G is k-colourable if and only if the profile of AF's we constructed can be rationalised using at most k values. □

This is bad news. But are there special cases where rationalisability is tractable after all? In the remainder of this section, we are going to explore this question. First, note that restricting attention to *complete preferences* does not help. Indeed, in the rationalisability problem used in our reduction, all agents already have preference orders that are complete (for each agent, we only had to specify her preferences with respect to the two values used to label the two arguments she is aware of).

Second, recall that Graph Colouring is *not* NP-hard for $k = 2$ colours, so our proof of intractability does not cover the case of exactly *two values*. Whether Proposition 9 can be strengthened to a bound of $k = 2$ or whether rationalisation with two values is tractable is an interesting open question.

Third, recall that our proof of Proposition 9 very heavily relies on the fact that different agents may be aware of different sets of arguments. This often is a reasonable assumption (Coste-Marquis et al., 2007), but the special case where all agents consider the exact *same set of arguments* certainly is also of interest. Whether deciding rationalisability for this special case is tractable (say, for the case of complete preferences, where the corresponding single-agent problem is tractable) is yet another interesting open question.

However, if we combine the two restrictions just considered, we are able to obtain a positive result. That is, if all agents are aware of the exact same set of arguments *and* if we are allowed to use at most two values, then deciding rationalisability is tractable:

Proposition 10 (Two Values and Common Argument Sets)**.** *Whether a profile of AF's over a common set of arguments can be rationalised by an AVAF with a given fixed master attack-relation and using at most two values can be decided in polynomial time.*

Proof. Let $\boldsymbol{AF} = (\langle Arg, \rightrightarrows_1 \rangle, \ldots, \langle Arg, \rightrightarrows_n \rangle)$ be a profile of AF's over the common set of arguments *Arg*, let $\rightharpoonup$ be the master attack-relation we are given, and let v_1 and v_2 be the values we are allowed to use for rationalisation. We begin with three observations that allow us to restrict the range of rationalisability problems we need to be able to handle. First, whether an agent is indifferent between v_1 and v_2 or whether she considers them incomparable has the same effect. So, w.l.o.g., we may restrict attention to complete preferences. Second, in case the profile is unanimous, i.e., in case the n defeat-relations are all the same, we are left with a single-agent problem, which by Proposition 6 can be solved in polynomial time. So, w.l.o.g., we may assume that the profile of AF's is not unanimous and thus—as every agent reports the same set of arguments—that the profile of preference orders also is not unanimous. Third, as always, rationalisation is only possible in case $(\rightrightarrows_i) \subseteq (\rightharpoonup)$ for all $i \in \mathcal{N}$. Furthermore, in case $(\rightrightarrows_i) = (\rightharpoonup)$, we can assume $v_1 \sim_i v_2$ and consider the rationalisation problem for the remaining agents independently. At the same time, rationalisation with $v_1 \sim_i v_2$ is impossible when $(\rightrightarrows_i) \subset (\rightharpoonup)$. So, w.l.o.g., we may assume that no agent fully agrees with the master attack-relation and that for every agent i we have either $v_1 \succ_i v_2$ or $v_2 \succ_i v_1$.

We are going to reduce our rationalisability problem to the following tractable variant of Graph Colouring with two colours: We are given an undirected graph $G = \langle V, E^{\doteq}, E^{=} \rangle$ with two kinds of edges, and ask whether it is possible to colour the vertices in V using only two colours in such a way that no two vertices with the same colour are linked by an edge in $E^{\doteq}$ and no two vertices with distinct colours are linked by an edge in $E^{=}$. This problem can be solved in polynomial time, because we can reduce it to an instance of the standard Graph Colouring problem by replacing each edge $(x, y) \in E^{=}$ with a new dummy vertex z and two normal edges $(x, z), (z, y) \in E^{\doteq}$.

We now build such a graph for our given rationalisability instance, with *Arg* being the set of vertexes. For any two arguments A and B, if there exists an agent i such that $A \rightharpoonup B$ but not $A \rightrightarrows_i B$, then A and B must get labelled with different values, so we assert $(A, B) \in E^{\doteq}$. On the other hand, if $A \rightharpoonup B$ and $A \rightrightarrows_i B$ for all $i \in \mathcal{N}$, we cannot immediately infer that A and B must get labelled with the same value, but only that every agent must like the value attached to A at least as much as the value attached to B. However, together with our assumptions that all agents have strict preferences and that the preference profile is not unanimous (i.e., that at least one agent prefers v_1 to v_2 and at least one agent prefers v_2 to v_1), we can make this inference and thus assert $(A, B) \in E^{=}$. By construction, the given profile is rationalisable under the given constraints if and only if the graph we have built can be coloured using only two colours, so we are done. □

As Graph Colouring is intractable for more than two colours, this construction cannot be generalised to higher bounds on the number of values. We also would like to emphasise that our proof exploits the fact that every agent is assumed to be aware of the same set of arguments. Without this assumption, it would be wrong to claim that absence of unanimity of the profile of AF's implies absence of unanimity of the preference profile. This in turn means that we cannot reduce the rationalisability problem to a problem with equality-

and inequality-constraints, but instead require more complex constraints to reason about the preferences of the agents. As previously mentioned, this leaves some room between Proposition 10 (tractability for two values and common arguments) and Proposition 9 (intractability for $k \geqslant 3$ values and arbitrary arguments) and the complexity of rationalisability when we impose only one of these two restrictions is unknown.

Proposition 10 suggests that rationalisability becomes easier when the argument sets the agents report are all exactly the same. The following simple observation shows that the opposite is also true: if the argument sets are all radically different from each other, then rationalisation also becomes easy.

Fact 11 (Mutually Exclusive Argument Sets)**.** *Any profile of AF's of the form* $\boldsymbol{AF} = (\langle Arg_1, \rightrightarrows_1 \rangle, \ldots, \langle Arg_n, \rightrightarrows_n \rangle)$*, with* $Arg_i \cap Arg_j = \varnothing$ *for all pairs of agents* $i, j \in \mathcal{N}$*, can be rationalised using just a single value.*

Proof. Observe that for $|Val| = 1$ there is only a single value-labelling and that therefore the preference orders are irrelevant. The master attack-relation $(\rightharpoonup) = (\rightrightarrows_1) \cup \cdots \cup (\rightrightarrows_n)$ achieves rationalisation in this case, as it ensures $(\rightrightarrows_i) = (\rightharpoonup)\restriction_{Arg_i}$ for all $i \in \mathcal{N}$. □

If all agents report mutually disjoint sets of arguments and we are given a fixed master attack-relation, we might require more than just one value to achieve rationalisation. Note that this case is covered by Proposition 8.

Recall that in case there is no bound on the number of values (or, equivalently, if $k = |Arg|$), we already know that rationalisation is tractable (as this follows from Proposition 8). Our final result in this section shows that the rationalisability problem remains tractable when the bound k is 'large'—in the sense of only reducing the number of allowed values by a constant d (relative to the maximum $k = |Arg|$).

Proposition 12 (Large Bound on Values)**.** *Let* $d \in \mathbb{N}$ *be an arbitrary constant. Whether a profile* $\boldsymbol{AF} = (\langle Arg_1, \rightrightarrows_1 \rangle, \ldots, \langle Arg_n, \rightrightarrows_n \rangle)$ *is rationalisable by an AVAF with a given fixed master attack-relation and at most* $k := |Arg_1 \cup \cdots \cup Arg_n| - d$ *values can be decided in polynomial time.*

Proof. Let $m := |Arg_1 \cup \cdots \cup Arg_n|$. There are $p := \binom{\binom{m}{2}}{d}$ ways of selecting d pairs from amongst all pairs of distinct arguments. This number is exponential only in d (not in m). Thus, as d is constant, p is polynomial. Note that p is a (generous) upper bound on the number of ways in which we can divide the m arguments into $k = m - d$ clusters: for any desired division into k clusters, there exists a choice of d pairs such that we obtain that clustering by merging exactly those pairs.

Note that it is not important *which* value is used to label a given argument: if rationalisation is possible at all, then it remains possible after any given permutation of the values. The class of all clusterings with k clusters thus represents all relevant value-labellings with k values. Also note that, if rationalisation is possible with fewer than k values, then it certainly is possible with exactly k values. So we only need to check labellings with exactly k values.

To summarise, we have shown that our original rationalisation problem can be reduced to polynomially many (namely, p) new rationalisation problems—each of them for the same fixed master attack-relation and its own fixed value-labelling. But each of these individual problems is polynomial by Proposition 8 (item d), so we are done. □

6. Rationalisability under Expansion Semantics

Recall that in Section 2 (in the discussion following Definition 4), we briefly mentioned an alternative approach to defining the rationalisability problem. The basic idea was that, rather than assuming that agents first become aware of a subset of the set of all alternatives and then reduce the master-attack relation restricted to that subset in line with their own individual preferences, they might instead first reduce the master attack-relation according to their preferences and only then choose which subset of arguments to report. In this section we explore this idea further. In the remainder of the paper, we shall refer to the concept of rationalisability fixed by Definition 4 as *rationalisability under the standard semantics* and we shall contrast that semantics with the alternative semantics to be defined next.

We require some additional terminology. Given two argumentation frameworks $\langle Arg, \rightrightarrows \rangle$ and $\langle Arg^c, \rightrightarrows^c \rangle$, we call the latter an *expansion* of the former, if it is the case that $Arg \subseteq Arg^c$ and $(\rightrightarrows) = (\rightrightarrows^c)\restriction_{Arg}$. The expansion of a profile of AF's then is defined accordingly: $(\langle Arg_1^c, \rightrightarrows_1^c \rangle, \ldots, \langle Arg_n^c, \rightrightarrows_n^c \rangle)$ is an expansion of $(\langle Arg_1, \rightrightarrows_1 \rangle, \ldots, \langle Arg_n, \rightrightarrows_n \rangle)$ if, for every agent $i \in \mathcal{N}$, it is the case that $\langle Arg_i^c, \rightrightarrows_i^c \rangle$ is an expansion of $\langle Arg_i, \rightrightarrows_i \rangle$.

Definition 5 (Rationalisability under Expansion Semantics)**.** *A profile of AF's, $\boldsymbol{AF} = (\langle Arg_1, \rightrightarrows_1 \rangle, \ldots, \langle Arg_n, \rightrightarrows_n \rangle)$, is called rationalisable under the expansion semantics for a given set of constraints, if there exist an expansion $\boldsymbol{AF}^c = (\langle Arg, \rightrightarrows_1^c \rangle, \ldots, \langle Arg, \rightrightarrows_n^c \rangle)$ of $\boldsymbol{AF}$ with $Arg = Arg_1 \cup \cdots \cup Arg_n$, an attack-relation $\rightharpoonup$ on Arg, a set of values Val with a mapping $val : Arg \to Val$, and a profile $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$ of preference orders on Val, all meeting said constraints, such that, for all agents $i \in \mathcal{N}$ and all arguments $A, B \in Arg$, it is the case that $A \rightrightarrows_i^c B$ if and only if $A \rightharpoonup B$ but not $val(B) \succ_i val(A)$.*

In other words, $\boldsymbol{AF}$ is rationalisable under the expansion semantics, if we can find a profile $\boldsymbol{AF}^c$ of AF's that has the following three properties:

- each agent i reports the same set of arguments, namely $Arg = Arg_1 \cup \cdots \cup Arg_n$;
- $\boldsymbol{AF}^c$ is an expansion of $\boldsymbol{AF}$, i.e., $(\rightrightarrows_i) = (\rightrightarrows_i^c)\restriction_{Arg_i}$ for all $i \in \mathcal{N}$;
- $\boldsymbol{AF}^c$ is rationalisable under the standard semantics.

Observe that when all agents agree on the set of arguments, i.e., if $Arg_i = Arg_j$ for all $i, j \in \mathcal{N}$, then the standard semantics and the expansion semantics coincide.

On top of the various types of constraints considered so far, we may also impose constraints on the kinds of expansions we permit. We consider three such constraints here:

- We say that the profile $\boldsymbol{AF}$ can be rationalised under the expansion semantics using a *maximal expansion*, if—on top of the conditions stated in Definition 5—we have $(\rightrightarrows_i^c) = (\rightrightarrows_i) \cup [(\rightharpoonup) \cap (\overline{Arg_i} \times Arg \cup Arg \times \overline{Arg_i})]$ for all $i \in \mathcal{N}$, where $\overline{Arg_i} = Arg \setminus Arg_i$.

- We say that the profile $\boldsymbol{AF}$ can be rationalised under the expansion semantics using a *minimal expansion*, if—on top of the conditions stated in Definition 5—we have $(\rightrightarrows_i^c) = (\rightrightarrows_i)$ for all $i \in \mathcal{N}$.

- We say that the profile $\boldsymbol{AF}$ can be rationalised under the expansion semantics using a *disjoint-value expansion*, if—on top of the conditions stated in Definition 5—we have $\{val(A) \mid A \in Arg_i\} \cap \{val(A) \mid A \notin Arg_i\} = \varnothing$ for all $i \in \mathcal{N}$.

Thus, when using a maximal expansion, we assume that every agent accepts *all* of the attacks involving (one or two) arguments she does not report, while when using a minimal expansion, we assume that every agent accepts *none* of those attacks. We may think of this as two possible default assumptions of how an agent deals with attacks she does not explicitly report, either because she is not aware of them or because she consciously chooses not to mention them: she either accepts all of those other arguments (maximal expansion) or she rejects them all (minimal expansion). When we do not impose either a maximality or a minimality constraint, then we are free to choose which of those 'virtual' attacks to include in the expansion. The disjoint-value expansion is intended to model the idea that the reason that an agent does not report a given argument is that she considers the value attached to that argument a 'taboo'. Thus, the values assigned to the arguments she does report and the values assigned to those she does not report cannot overlap.

In principle, the full research agenda of analysing the rationalisability of an observed profile of AF's under different kinds of constraints, as initiated in this paper for the standard semantics, may also be carried out for the expansion semantics. Rather than attempting to do so here, we restrict ourselves to a small sample of results that illustrate the relationship between the standard semantics and the expansion semantics. We begin by showing that, in the absence of any constraints pertaining to the kind of expansion sought, the expansion semantics in fact coincides with the standard semantics.

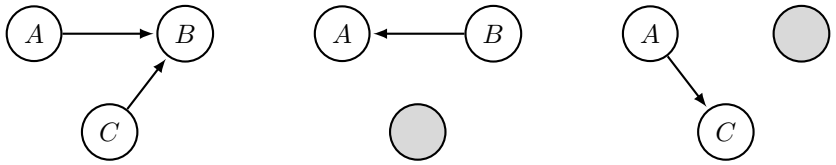
Proposition 13 (Standard and Expansion Semantics)**.** *Rationalising a given profile of AF's for a given set of constraints is possible under the standard semantics if and only if doing so is possible under the expansion semantics.*

Proof. For the left-to-right direction, suppose we have found a rationalisation of a given profile $\boldsymbol{AF} = (\langle Arg_1, \rightrightarrows_1 \rangle, \ldots, \langle Arg_n, \rightrightarrows_n \rangle)$ under the standard semantics. Note that, by Definition 4, such a rationalisation in fact involves agent preferences that range over the full set of values, even those values assigned only to arguments a given agent is not aware of. Now, for each agent $i \in \mathcal{N}$, define $\rightrightarrows_i^c$ as the result of reducing the union of $\rightrightarrows_i$ and $(\rightharpoonup)\restriction_{Arg \setminus Arg_i}$ with $\succcurlyeq_i$, where $\rightharpoonup$ is the master attack-relation found under the standard semantics, Arg is the set of all arguments, and $\succcurlyeq_i$ is agent i's preference order found under the standard semantics. It is now easy to check that this constitutes a valid rationalisation under the expansion semantics.

For the other direction, suppose that, under the expansion semantics, we have found a way to rationalise $\boldsymbol{AF}$, and suppose we have used some specific expansion $\boldsymbol{AF}^c = (\langle Arg, \rightrightarrows_1^c \rangle, \ldots, \langle Arg, \rightrightarrows_n^c \rangle)$ to do so. Now restrict attention to how this rationalisation acts on each $\langle Arg_i, \rightrightarrows_i \rangle$, keeping in mind that $\rightrightarrows_i$ and $\rightrightarrows_i^c$ coincide on Arg_i. As inspection of Definitions 4 and 5 confirms, in this manner we obtain precisely the conditions of rationalisability under the standard semantics. □

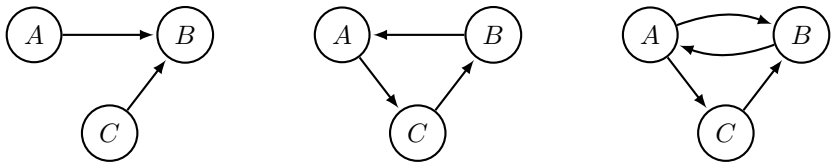
Importantly, in the proof of the left-to-right direction of Proposition 13 we made use of the assumption that there are no constraints on the type of expansion required (because the set of constraints imposed is the same under both semantics). Once we impose such constraints, the expansion semantics (potentially) becomes more restrictive. We are now going to illustrate this point with several results regarding the case of maximal expansions. We start with an example that shows that when we require complete preferences and maximal expansions, then the standard semantics and the expansion semantics differ.

Example 4. *Consider a profile of three agents, with* $Arg_1 = \{A, B, C\}$ *and both* $A \rightrightarrows_1 B$ *and* $C \rightrightarrows_1 B$, $Arg_2 = \{A, B\}$ *and* $B \rightrightarrows_2 A$, *and* $Arg_3 = \{A, C\}$ *and* $A \rightrightarrows_3 C$. *This profile is shown below, with the arguments a given agent does not report being represented in grey:*



Now suppose we want to rationalise this profile using complete preferences and suppose all other constraints are as favourable as possible, i.e., we are allowed to use three values, $\{v_A, v_B, v_C\}$, *and the master attack-relation* $\rightharpoonup$ *to be used is exactly the union of the three individual defeat-relations, i.e.,* $(\rightharpoonup) = \{(A, B), (B, A), (A, C), (C, B)\}$. *Fix* $val(A) = v_A$, $val(B) = v_B$, *and* $val(C) = v_C$. *The following complete preferences achieve rationalisation under the standard semantics:* $v_C \succ_1 v_A \succ_1 v_B$, $v_B \succ_2 v_A$, *and* $v_A \succ_3 v_C$.

Using the expansion semantics with the maximal expansion, on the other hand, rationalisation is not possible. To see this, note first that the completed profile $\boldsymbol{AF}^c$ *we would have to rationalise in this case is the following:* $A \rightrightarrows^c_1 B$, $C \rightrightarrows^c_1 B$; $B \rightrightarrows^c_2 A$, $A \rightrightarrows^c_2 C$, $C \rightrightarrows^c_2 B$; *and* $A \rightrightarrows^c_3 C$, $C \rightrightarrows^c_3 B$, $A \rightrightarrows^c_3 B$, $B \rightrightarrows^c_3 A$. *This is shown below:*



But this profile is not rationalisable using complete preferences: For agent 2, we certainly have to have $v_B \succ_2 v_A$ *in order to remove the attack from A to B. Then, if we choose* $v_C \succ_2 v_A$, *the attack from A to C would get removed, while if choose* $v_A \succcurlyeq_2 v_C$, *we would get* $v_B \succ v_C$ *by transitivity and thus the attack from C to B would get removed.*

This example notwithstanding, we are now going to see that in the absence of strong constraints, any profile that can be rationalised under the standard semantics can also be rationalised under the expansion semantics with maximal expansions. The basic intuition is that, if we have a sufficiently large number of values at our disposal, we can set things up in such a way that each agent is indifferent between the values of any of the arguments she did not report. We now turn this intuition into a precise result for the case of a fixed master attack-relation. Recall that, for the standard semantics, this case is covered by part (b) of Proposition 8, in conjunction with Proposition 2.

Proposition 14 (Maximal Expansions)**.** *Rationalising a profile of AF's by an AVAF with a given fixed master attack-relation is possible under the standard semantics if and only if doing so is possible under the expansion semantics using a maximal expansion.*

Proof. The right-to-left direction follows from Proposition 13. For the left-to-right direction, assume the given AF is rationalisable under the standard semantics. Recall from the proof of Proposition 2 that this means that there also must be a rationalisation under which every argument is assigned its own value. Now, for each agent $i \in \mathcal{N}$, take the preference order $\succcurlyeq_i$

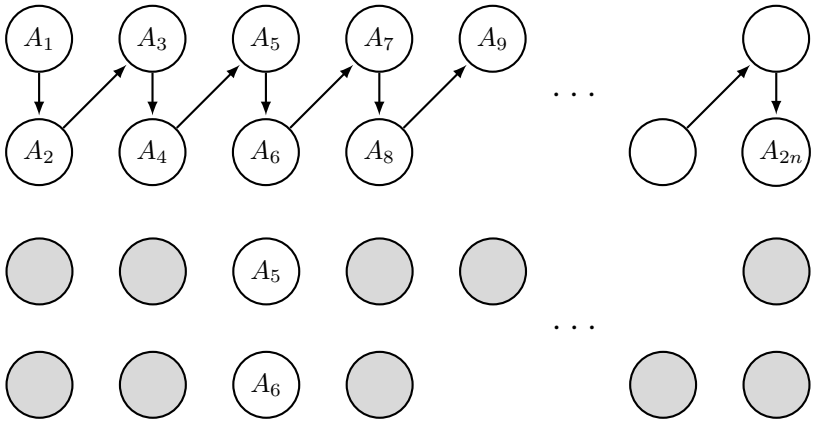
used for the latter rationalisation as far as the values corresponding to the arguments in Arg_i are concerned, and complete it to a preference order over the full set of values, corresponding to Arg, by declaring any of the new pairs of values preferentially incomparable. This achieves the required rationalisation. □

We stress that the construction used in the proof of Proposition 14 involves incomplete preferences. If we add the constraint that preferences must be complete, then the proof does not go through any longer and the corresponding claim would be false (as is demonstrated by Example 4).

Recall that Proposition 14 does not involve any constraints on the number of values to be used. Indeed, we may require more values to rationalise a profile under the expansion semantics using the maximal expansion than we require to rationalise the same profile under the standard semantics. The following result shows that the ratio between the numbers of values required can grow arbitrarily large.

Proposition 15 (Value Ratio for Maximal expansions)**.** *The ratio between the number of values required to rationalise a given profile of AF's by an AVAF with a given fixed master attack-relation under the expansion semantics using a maximal expansion and the number of values required to do the same under the standard semantics, in the worst case, cannot be bounded from above by a constant.*

Proof. To prove the claim, we will show how to construct a family of profiles of AF's and a master attack-relation with $2n$ arguments each, where n as usual is the number of agents, such that, for each profile in that family, rationalisation under the standard semantics is possible with two values, while rationalisation under the expansion semantics with maximal expansions requires $\Omega(\sqrt{n})$ (or, equivalently, $\Omega(\sqrt{|Arg|})$) values. Let $Arg = \{A_1, A_2, \ldots, A_{2n}\}$ and fix a master attack-relation as follows: $(\rightharpoonup) = \{(A_k, A_{k+1}) \mid k < 2n\}$. Suppose each agent $i \in \mathcal{N}$ is only aware of the arguments A_{2i-1} and A_{2i} and reports an empty defeat-relation, i.e., $Arg_i = \{A_{2i-1}, A_{2i}\}$ and $(\rightharpoonup_i) = \varnothing$. We visualise this scenario by indicating the master attack-relation and the AF reported by agent 3 below:



Under the standard semantics, rationalisation is possible with just two values: simply set $val(A_{2k-1}) = v_1$ and $val(A_{2k}) = v_2$ for all $k < 2n$, and choose the preference order $\succcurlyeq_i$ with $v_2 \succ_i v_1$ for every agent $i \in \mathcal{N}$. On the other hand, under the expansion semantics with maximal expansions, we would have to rationalise a completed profile for which each agent i

is only missing the edge from A_{2i-1} to A_{2i} from the master attack-relation. Thus, for every $i \in \mathcal{N}$ we would have to require $val(A_{2i}) \succ_i val(A_{2i-1})$ but also $val(A_{k+1}) \not\succ_i val(A_k)$ for every $k < 2n$ except $k = 2i - 1$. Due to the first condition, we must label every ordered pair of 'consecutive' arguments (A_{2i-1}, A_{2i}) with distinct values. But due to the second condition, we cannot use the same ordered pair of values for any two such ordered pairs of arguments (i.e., $val(A_{2i-1}) \neq val(A_{2j-1})$ or $val(A_{2i}) \neq val(A_{2j})$ whenever $i \neq j$). With $|Val|$ values at our disposal, we can create $|Val| \cdot (|Val| - 1)$ ordered pairs of distinct values, i.e., we must have $|Val| \cdot (|Val| - 1) \geqslant |Arg| = 2n$ and thus $|Val| \in \Omega(\sqrt{n})$. □

To conclude this section, let us contrast our findings for the case of maximal expansions with the two other types of expansions we have defined.

For disjoint-value expansions, it is easy to derive results similar to Propositions 14 and 15. We shall restrict ourselves to informal statements. First, any profile of AF's that is rationalisable under the standard semantics for a fixed master attack-relation is also rationalisable under the expansion semantics with a disjoint-value expansion. The proof is the same as for Proposition 14, except that there is no need to alter the agents' preferences over the values corresponding to arguments they are not aware of. Thus, the same result also holds if we require preferences to be complete, which was not the case for maximal expansions. Second, the number of values required for rationalisation under the expansion semantics with a disjoint-value expansion can be arbitrarily higher than the number required under the standard semantics. To prove this, it suffices to consider scenarios where we have one argument A_S for every nonempty subset S of the set of agents $\mathcal{N}$, with exactly the agents in S being aware of A_S. Rationalising such a profile using disjoint-value expansions—if possible at all—requires one new value for every single argument. On the other hand, under the standard semantics, a single value will be enough in some cases, e.g., when every agent reports exactly the master attack-relation, restricted to the set of arguments she is aware of.

For minimal expansions, interestingly, as the following example demonstrates, we cannot obtain a result analogous to Proposition 14. Thus, this constraint makes the expansion semantics significantly more demanding than the standard semantics.

Example 5. *Consider again the profile of AF's discussed in the context of Example 4, but now assume that the agents use the minimal expansions. Then, after expansion, agent 3 will have the AF* $\langle Arg_3^c, \rightrightarrows_3^c \rangle$ *with* $Arg_3^c = \{A, B, C\}$ *and* $A \rightrightarrows_3^c C$*. Her preferences over the values assigned to the arguments have to allow her to remove the attack from A to B as well as the one from B to A. But this is impossible, even with incomplete preferences and an arbitrary number of values.*

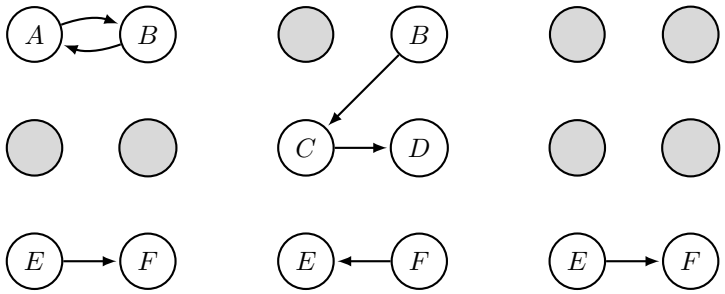
7. Application Scenarios

There are a number of different application scenarios where dealing with questions of rationalisability will be valuable. In this section, we list and illustrate some of them.

First, given the growing interest in the abstract argumentation research community in questions of aggregation of AF's (Coste-Marquis et al., 2007; Tohmé et al., 2008; Bodanza & Auday, 2009; Dunne et al., 2012; Bodanza et al., 2017), it is important to have a clear understanding for what types of scenarios the question of aggregation is in fact relevant.

Our notion of rationalisability provides a suitable definition for this purpose. It allows for a systematic scan of the different examples used in the literature—not to dismiss those failing the test, but to point out that one must be careful with the interpretation used.

Example 6. *Let us see whether the example given by Coste-Marquis et al. (2007, Example 7) passes this test. We are given* $AF_1 = \langle\{A, B, E, F\}, \{(A, B), (B, A), (E, F)\}\rangle$, $AF_2 = \langle\{B, C, D, E, F\}, \{(B, C), (C, D), (F, E)\}\rangle$, *and* $AF_3 = \langle\{E, F\}, \{(E, F)\}\rangle$*:*



This profile indeed does pass the test under the standard semantics. It can be rationalised by using as master attack-relation the union of the individual relations. It is sufficient to set $v_E \neq v_F$*, while A, B, C, and D can all take the same value, either that of E or that of F. Thus, two values suffice.*

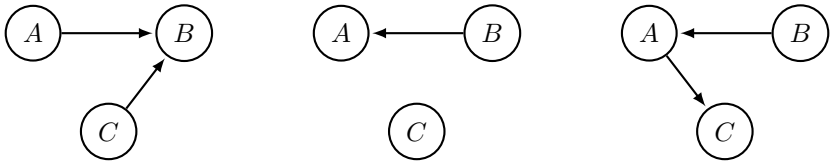
Under the minimal expansion semantics, we observe that the profile is not rationalisable, because for instance for the pair of arguments A and B, agent 3 would need to include at least one of the two attacks. On the other hand, when using the maximal expansion, two values suffice as in the standard case.

Under the disjoint-value expansion semantics, we have the following constraints, on top of $v_E \neq v_F$*:* $\{v_C, v_D\} \cap \{v_A, v_B, v_E, v_F\} = \emptyset$*, due to agent 1;* $v_A \notin \{v_B, v_C, v_D, v_E, v_F\}$*, due to agent 2; and* $\{v_A, v_B, v_C, v_D\} \cap \{v_E, v_F\} = \emptyset$*, due to agent 3. Hence, at least five values are required, as only C and D can share the same value. Five values indeed suffice, for instance by taking* $v_E \succ_1 v_F$ *and* $v_A \sim_1 v_B$*,* $v_F \succ_2 v_E$ *and* $v_B \sim_2 v_C$*, and* $v_E \succ_3 v_F$*.*

The second application of our work concerns aggregation itself. In a scenario where multiple AF's need to be aggregated, we may use the notion of rationalisability to choose between alternative aggregation techniques, depending on the result of the rationalisability test. For example, if a profile turns out to be rationalisable for a given preference model (e.g., for complete preference orders), we may reasonably assume that this model is a good abstraction of reality and aggregate the AF's by aggregating the inferred preferences (which is a much better studied problem than that of aggregating AF's). For instance, we may use the well-known Kemeny rule (Kemeny, 1959; Zwicker, 2016) to aggregate the preferences,[7] and then apply the collective preference order obtained to the master attack-relation inferred.

Example 7. *Consider the following profile of AF's, in which each of the three agents reports the same set of arguments* $\{A, B, C\}$*:*

7. For a given profile of preference orders, the Kemeny rule returns the preference order that minimises the number of times an agent disagrees on the relative ranking of two alternatives. In other words, the Kemeny rule minimises the sum of the Kendall tau distances between the outcome and the individual preference orders.



The union of the individual AF's together with the following preferences achieves rationalisation under the standard semantics: $v_C \succ_1 v_A \succ_1 v_B$, $v_B \succ_2 v_C \succ_2 v_A$, *and* $v_B \succ_3 v_A \succ_3 v_C$. *Observe that this is the only rationalisation with complete and strict preferences. The result of applying the Kemeny rule to this profile of preferences is* $v_B \succ v_C \succ v_A$ *(with a Kendall tau distance of 3 from the collection of preferences). Hence, the aggregated AF with this technique would be the same as the AF of agent 2.*

But when rationalisation fails, this approach does not make sense, and we should look for a different method of aggregation. In such a case, there is a more substantial disagreement between the agents: maybe the model of preferences has to be changed, maybe the agents differ on the assignment of values to arguments, or maybe the agents interpret the arguments differently. Importantly, failure of rationalisation can also provide hints as to where disagreement occurs.

A third application we foresee is in the context of online debating platforms, where value-based argumentation systems already are used as a modelling tool (Pulfrey-Taylor, Henthorn, Atkinson, Wyner, & Bench-Capon, 2011). In this context, AF's are (typically) not obtained via a one-shot process, but rather retrieved interactively, by monitoring the utterances of the participants. Our approach could be used to detect inconsistencies as they occur, and thus to trigger clarification questions on the fly.

Example 8. *Suppose the following sequence of utterances occurs in a given debate:*

- *Agent 1: A defeats B.*
- *Agent 2: B defeats A.*
- *Agent 3: There is no defeat between A and B.*

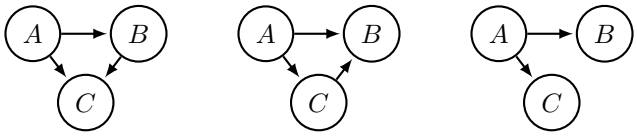
At this stage it is clear that this collection of AF's cannot be rationalised, because agent 3 would have to both prefer the value of A over that of B, and the value of B over that of A. A clarification is required to identify the mismatch. For example, the system could ask agent 3 whether she really believes there is no attack between A and B.

Interestingly, a similar dynamic perspective to solve inconsistencies in a framework mixing opinion polling and argumentation has recently been proposed by Rago and Toni (2017), albeit in their case the problem is to rationalise the *votes* of users. Interleaving the elicitation of preferences over values within a dialectical process is also proposed by Bench-Capon et al. (2007). But these authors take a different perspective. While we assume that an agent's preferences over values are fixed from the outset and just need to be 'discovered' during rationalisation, they do not take this assumption for granted. Instead, the ranking of values is built as part of the dialectical process, whereby an agent (in their case, just one agent) aims at rationalising her position (i.e., the arguments she wants to see accepted or rejected).

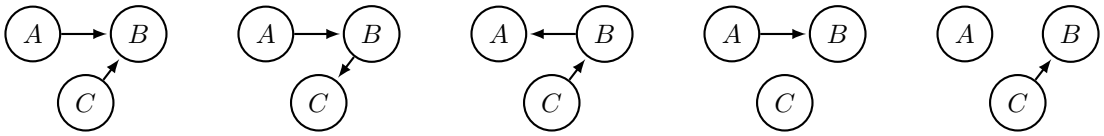
This leads us naturally to our final point of discussion, which concerns the nature of what is observed. So far we have assumed that the agents express AF's, which we can observe

directly. But in many situations, it may be more natural to assume that each agent only reports the set of arguments she accepts (a so-called *extension*), or a (partial) *labelling* of the arguments with the labels 'accept' and 'reject'. Dunne et al. (2014) have addressed the challenging problem of inferring an AF from such an extension (or a set of such extensions) that could serve as an explanation for the behaviour observed. Of course, there often will be *many* possible AF's that could explain a given set of accepted arguments. Our approach could be used to narrow down the range of possible explanations when performing this task for several agents in parallel, by imposing the constraint that the profile of AF's we infer, one for each extension observed, should be rationalisable.

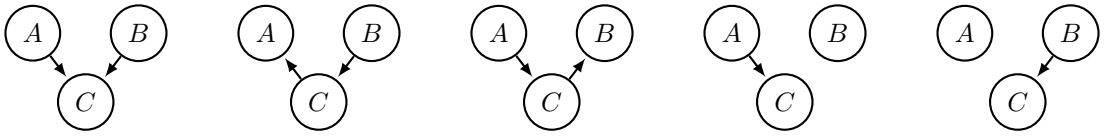
Example 9. *Suppose there are three agents and three arguments. Agent 1 reports extension* $\{A\}$*, agent 2 reports extension* $\{A, C\}$*, and agent 3 reports extension* $\{A, B\}$*. Suppose these reports have come about by means of each agent applying one of the well-known semantics proposed by Dung (1995) to some AF declared over the full set* $\{A, B, C\}$*.*[8] *For the sake of simplicity, let us exclude the possibility of mutual attacks between pairs of arguments. Then agent 1, who only considers A acceptable, must have one of the following three AF's:*



Now, for agent 2, there must be no attack between A and C, while B must get attacked by at least one of them. This leaves five possible cases:



Finally, for agent 3, we can have no attack between A and B. She must have generated her position on the basis of one of the following five AF's:



But now, in terms of rationalisation, we see that some combinations are impossible, such as for instance the profile consisting of the third AF for agent 1, the second for agent 2, and the third for agent 3. To see this, note that the master attack-relation would have to contain both $B \rightharpoonup C$ *(for agent 2) and* $C \rightharpoonup B$ *(for agent 3). But then agent 1 would have to have one of these attack relations in her system, as she cannot both strictly prefer the value of B to that of C and vice versa. While this does not allow us to uniquely define a profile of AF's, this method can nevertheless guide the search amongst the AF's that are compatible with the extensions observed.*

8. The details of the formal definition of this kind of semantics are not important for our example. In a nutshell, what matters here is that you cannot accept two arguments such that one attacks the other, and if you accept an argument that is attacked by one of those you reject, then you also must accept an argument that attacks that attacker. We further assume that you accept all unattacked arguments.

Similar ideas may also have useful applications in the context of analysing people's decisions *a posteriori*. A recent example for an application of this kind is the analysis of a participatory decision setting involving an environmental project in Québec, which was carried out by Tremblay and Abi-Zeid (2016). In their work, they first extracted an AF with 20 arguments, labelled by 7 values, from the debates they analysed. They then imposed a number of technical constraints, eventually obtaining 18 subgraphs of the master AF as possible candidates for the kind of AF that may in practice have guided the deliberations of the committee responsible for taking a decision about which arguments to accept. They then analysed each of these 18 AF's in combination with one of the possible preference orders over the 7 values, to test whether and how often the decision recommended by a given AF coincides with the decision actually observed in practice. (That decision consisted in accepting 5 of the 20 arguments considered.) To make this analysis manageable, the choice of the 18 AF's considered required a number of judgment calls. Here, the concept of rationalisability may offer an alternative route. For a set of arguments we observe to have been accepted in practice, we may first induce a number of possible AF's that could explain this extension, using the approach of Dunne et al. (2014), and then apply our rationalisation approach to check whether any of these AF's is rationalisable, given the constraints regarding values we have been able to extract from the debate.

8. Conclusion

We have introduced the concept of *rationalisability* of a profile of abstract argumentation frameworks, proposed a specific instantiation of the general idea in terms of social values associated with the arguments and preferences over those values held by the agents, and studied the resulting decision problem from an algorithmic point of view, for several types of constraints on admissible solutions and two possible interpretations for the fact that different agents may report different sets of arguments. We have been able to show that the single-agent rationalisability problem is tractable for all the constraints considered. These positive results extend to the multiagent case for several types of constraints. However, in the presence of a constraint limiting the number of values we may use, the most general variant of the multiagent problem is NP-complete.[9] Finally, we have discussed possible application scenarios where the notion of rationalisability may play a role.

While our technical results offer a good initial overview of the landscape of rationalisability, our work also pinpoints a number of interesting open questions. Let us briefly recall the three main technical questions we have left open. The first concerns the complexity of single-agent rationalisability with a limited number of values for possibly incomplete preferences: Does Proposition 6, which establishes tractability of this problem under the additional assumption that preferences are complete, continue to hold when we drop the completeness constraint? The second and the third open question both concern the multiagent rationalisability problem with a limited number of values: Does Proposition 9, which establishes intractability of this problem for bounds $k \geqslant 3$ on the number of values, continue

9. Recall that we have assumed attack relations to be irreflexive. The complexity of the rationalisability problem is not affected by this assumption. As no assignment of values and choice of preference orders can ever cancel out a self-attack, all you need to do on top of checking our existing conditions is checking that all agents agree on all self-attacks.

to hold for a bound of $k = 2$? And does it continue to hold in case we restrict attention to cases where all agents report the same set of arguments? We have been able to show that the problem becomes tractable in case *both* of these restrictions are imposed together, but it remains unclear how each of them alone affects complexity. Settling any of these questions would constitute a valuable contribution to this area of research.

Besides addressing these concrete questions, future work might also be directed towards identifying and analysing further constraints on rationalisation in general (besides, e.g., bounds on the number of values), further constraints on admissible expansions under the expansion semantics (besides, e.g., the disjoint-valuedness constraint inspired by the notion of 'taboo'), and further restrictions of the general framework (besides, e.g., the restriction to a common argument set).

Future work should also investigate alternative instantiations of the general idea of rationalisability expounded here. For instance, as mentioned already in the introduction, the model of Bench-Capon (2003) is but one approach to modelling the emergence of different individual argumentation frameworks. Defining the rationalisability problem for competing approaches is likely to be fruitful as well. Finally, it is important to keep in mind that Dung's model of abstract argumentation is just that: an *abstract* model of argumentation. Other formalisms, which also model the internal structure of arguments, come closer to real forms of argumentation occurring between people. Therefore, applying our approach to such richer models of argumentation also is an important direction for future work.

Acknowledgments

This paper extends a paper presented at AAMAS-2016 (Airiau, Bonzon, Endriss, Maudet, & Rossit, 2016). We are grateful for the feedback received from several anonymous reviewers. Our work was partly supported by COST Action IC1205 on Computational Social Choice and by project AMANDE ANR-13-BS02-0004 of the French National Research Agency.

References

Airiau, S., Bonzon, E., Endriss, U., Maudet, N., & Rossit, J. (2016). Rationalisation of profiles of abstract argumentation frameworks. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2016)*, pp. 350–357. IFAAMAS.

Amgoud, L., & Cayrol, C. (2002). A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, *34*, 197–216.

Amgoud, L., Dimopoulos, Y., & Moraitis, P. (2008). Making decisions through preference-based argumentation. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR-2008)*, pp. 113–123.

Amgoud, L., & Vesic, S. (2011). A new approach for preference-based argumentation frameworks. *Annals of Mathematics and Artifcial Intelligence*, *63*(2), 149–183.

Atkinson, K., & Bench-Capon, T. J. M. (2016). Value based reasoning and the actions of others. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI-2016)*. IOS Press.

Baumann, R. (2012). What does it take to enforce an argument? Minimal change in abstract argumentation. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012)*. IOS Press.

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

Bench-Capon, T. J. M., Doutre, S., & Dunne, P. E. (2007). Audiences in argumentation frameworks. *Artificial Intelligence*, *171*(1), 42–71.

Black, E., & Hunter, A. (2012). A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, *22*(1), 55–78.

Bodanza, G. A., & Auday, M. R. (2009). Social argument justification: Some mechanisms and conditions for their coincidence. In *Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-2009)*. Springer-Verlag.

Bodanza, G. M., Tohmé, F. A., & Auday, M. R. (2017). Collective argumentation: A survey of aggregation issues around argumentation frameworks. *Argument & Computation*, *8*(1), 1–34.

Booth, R., Kaci, S., & Rienstra, T. (2013). Property-based preferences in abstract argumentation. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT-2013)*. Springer-Verlag.

Caminada, M., & Pigozzi, G. (2011). On judgment aggregation in abstract argumentation. *Journal of Autonomous Agents and Multiagent Systems*, *22*(1), 64–102.

Cartwright, D., & Atkinson, K. (2009). Using computational argumentation to support *e*-participation. *IEEE Intelligent Systems*, *24*(5), 42–52.

Cayrol, C., Doutre, S., Lagasquie-Schiex, M., & Mengin, J. (2002). "Minimal defence": A refinement of the preferred semantics for argumentation frameworks. In *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning (NMR-2002)*.

Coste-Marquis, S., Devred, C., Konieczny, S., Lagasquie-Schiex, M.-C., & Marquis, P. (2007). On the merging of Dung's argumentation systems. *Artificial Intelligence*, *171*(10–15), 730–753.

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

Dunne, P. E., Dvořák, W., Linsbichler, T., & Woltran, S. (2014). Characteristics of multiple viewpoints in abstract argumentation. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR-2014)*.

Dunne, P. E., Marquis, P., & Wooldridge, M. (2012). Argument aggregation: Basic axioms and complexity results. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA-2012)*. IOS Press.

Endriss, U., & Grandi, U. (2017). Graph aggregation. *Artificial Intelligence*, *245*, 86–114.

Gabbriellini, S., & Torroni, P. (2013). Arguments in social networks. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2013)*. IFAAMAS.

Grabmair, M., & Ashley, K. D. (2011). Facilitating case comparison using value judgments and intermediate legal concepts. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL-2011)*. ACM.

Greige, L. (2016). Rationalisation of argumentation profiles: An experimental study. Research internship report, LIP6, UPMC Université Paris 6.

Grossi, D., & van der Hoek, W. (2013). Audience-based uncertainty in abstract argument games. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI-2013)*.

Hochbaum, D. S., & Naor, J. (1994). Simple and fast algorithms for linear and integer programs with two variables per inequality. *SIAM Journal on Computing*, *23*(6), 1179–1192.

Kaci, S., & van der Torre, L. (2008). Preference-based argumentation: Arguments supporting multiple values. *International Journal of Approximate Reasoning*, *48*(3), 730–751.

Karp, R. M. (1972). Reducibility among combinatorial problems. In *Proceedings of a Symposium on the Complexity of Computer Computations*. Plenum Press.

Kemeny, J. (1959). Mathematics without numbers. *Daedalus*, *88*(4), 577–591.

Modgil, S. (2009). Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, *173*(9), 901–934.

Modgil, S., & Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, *195*, 361–397.

Pulfrey-Taylor, S., Henthorn, E., Atkinson, K., Wyner, A., & Bench-Capon, T. J. M. (2011). Populating an online consultation tool. In *Proceedings of the 24th Annual Conference on Legal Knowledge and Information Systems (JURIX-2011)*, pp. 150–154. IOS Press.

Rago, A., & Toni, F. (2017). Quantitative argumentation debates with votes for opinion polling. In *Proceedings of the 20th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA-2017)*. Springer-Verlag. To appear.

Rahwan, I., & Tohmé, F. A. (2010). Collective argument evaluation as judgement aggregation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*. IFAAMAS.

Roberts, F. S. (1979). *Measurement Theory*. Addison-Wesley.

Searle, J. R. (2001). *Rationality in Action*. MIT Press.

Tohmé, F. A., Bodanza, G. A., & Simari, G. R. (2008). Aggregation of attack relations: A social-choice theoretical analysis of defeasibility criteria. In *Proceedings of the 5th International Symposium on Foundations of Information and Knowledge Systems (FoIKS-2008)*. Springer-Verlag.

Tremblay, J., & Abi-Zeid, I. (2016). Value-based argumentation for policy decision analysis: Methodology and an exploratory case study of a hydroelectric project in Québec. *Annals of Operations Research*, *236*(1), 233–253.

Zwicker, W. S. (2016). Introduction to the theory of voting. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (Eds.), *Handbook of Computational Social Choice*, chap. 2. Cambridge University Press.