Preventing Manipulation in Aggregating Audiences in Value-Based Argumentation Frameworks

Grzegorz LISOWSKI [a], Sylvie DOUTRE [b], Umberto GRANDI [b]
[a] *University of Amsterdam*
[b] *Université Toulouse 1 Capitole, IRIT*

Abstract. In the context of value-based argumentation frameworks, and collective decisions, we are interested in the behavior of agents who are willing to misrepresent their sincere beliefs about the ranking of values in order to ensure that their desired decision is agreed upon. We study the application of preference aggregation mechanisms towards reaching a compromise ranking of values for a group of agents. Further, we study under which conditions agents can manipulate the outcome of the discussion. We investigate the computational complexity of this problem.

1. Introduction

Argumentation is an inherently multi-agent phenomenon. It often occurs when agents exchange information, aiming at reaching a collective view with respect to some issue. However, it is not clear how to conceptualize the multi-agent character of argumentation. One of the approaches towards solving this problem relies on the applications of *aggregation* methods, associated with social choice theory. In such approaches, a compromise structure of argumentation is provided for a number of agents, representing distinctive views on it. Further, when mechanisms merging agents' views about the aspects of argumentation are considered, agents might be willing to provide information conflicting with their beliefs, to obtain a better outcome for themselves.

Intuitively, the goal of collectively solving argumentation problems is to select the best arguments taking into account all relevant information that agents have at their disposal and to fairly combine views on the structure of argumentation. Agents can have preferences over accepted arguments, for instance if acceptance of some distinguished arguments is determining the choice of some *decision*. Then, they might decide not to submit arguments that they know about, as considered by Rahwan and Larson (2008). Also, they can misrepresent their views on the strength of arguments to ameliorate the outcome of discussion for themselves. We can also notice that arguments might appeal to particular values, which are of diversified importance to a selector of arguments. It is then plausible to assume that an attack on a strong argument from an argument of little importance should not be taken into account.

Several approaches towards capturing the differences in the strength of arguments have been introduced. One of them, *value-based argumentation*, was provided by Bench-

Capon (2003). In this framework it is assumed that arguments appeal to specific values, which are of a distinctive importance to a particular agent. Then, an attack can be blocked from her perspective if she ranks the value of the attacked argument higher than the value of its attacker. This approach is suited to the problem of argumentation-based decision-making, in which factors other than credibility of information are important while assessing the acceptability of an argument. Also, it provides a clear justification for the determined strength of arguments. This is important when an argumentation serves as a support for decision-making; justification of strength of arguments contributes to the justification of a decision. It is worth noting that this constitutes a strong advantage of this approach over assigning preferences over arguments directly, as in the preference-based argumentation (e.g Amgoud & Cayrol, 2002). Bench-Capon's approach makes sure that agents only consider some arguments as stronger than another, if they have a good reason to do so.

Recently, value-based argumentation has been studied in the multi-agent argumentation setting. Airiau, Bonzon, Endriss, Maudet, and Rossit (2016) investigated the problem of *rationalization* of disagreements between agents with respect to the structure of attacks between arguments. They propose to use value-based argumentation for this purpose. In the current paper, we are considering a related problem of finding an argumentation framework constituting a *compromise* between agents disagreeing about the structure of the attack relation because of their disagreement with respect to the importance of values. Following Pu, Luo, Zhang, and Luo (2013) we consider aggregation of agents' views on the importance of values with *preference aggregation* mechanisms. Then, we study the possibility of manipulation in aggregating agents' views on the strength of arguments, determined by the importance of values they appeal to. We will consider argumentation which results in a decision to either accept a certain action, or to reject it.

In Section 2 we present the basic framework in which the results are situated. We introduce value-based argumentation, acceptability semantics, and we present the way to employ it for group decision-making. Later, in Section 3, we study the ways of manipulating aggregation in the context of value-based argumentation frameworks. For the sake of simplicity our investigations are restricted to the grounded semantics. We finish with conclusions and suggestions for future work.

2. Preliminaries

Dung (1995) introduces argumentation frameworks. Such frameworks consist of a set of arguments and of a binary relation indicating which arguments are in conflict.

Definition 2.1 *An* argumentation framework *(AF) is a pair $AF = \langle A, \rightarrow \rangle$, where A is a set of arguments and $\rightarrow \subseteq A^2$ is the attack relation.*

An argumentation framework is then a directed graph, which nodes are the arguments, and edges represent the attacks between arguments.

A set of arguments S is said to defend an argument a if for any attacker of a there is some member of S which attacks it.

Definition 2.2 *Given an argumentation framework $AF = \langle A, \rightarrow \rangle$, a set of arguments $S \subseteq A$ and some argument $a \in A$, S* defends *a iff for any $b \in A$ such that $b \rightarrow a$ there is an $a' \in S$ such that $a' \rightarrow b$. We say that S defends a set of arguments $S' \subseteq A$ iff S defends all $a \in S'$.*

A function $F : 2^A \rightarrow 2^A$ assigns every $S \subseteq A$ the set of all arguments that S defends. Also, S is said to be self-defended *if S defends S.*

The notion of defense is used to determine when a set of arguments can be rationally selected as an outcome of a discussion. In addition to this notion, the following criteria for selecting sets of arguments have been considered:

Definition 2.3 *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework, and $S \subseteq A$. S is:*

- Conflict-free: *iff there are no $a, b \in S$ such that $a \rightarrow b$. We refer to the set of all conflict-free sets of AF as CFR_{AF}.*
- Admissible: *iff S is conflict-free and self-defended. We refer to the set of all admissible sets of AF as ADM_{AF}.*

Based on these criteria, acceptability semantics define which sets of arguments (extensions) can be collectively accepted (Dung, 1995).

Definition 2.4 *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework, and $S \subseteq A$. S is:*

- Complete: *iff S is admissible and $F(S) = S$. We refer to the set of all complete extensions of AF as CMP_{AF}.*
- Grounded: *iff S is the minimal complete extension of AF w.r.t. set inclusion. We refer to the grounded extension of AF as $GRND_{AF}$.*

It can be noticed that the grounded extension is always unique. For simplicity, it is this uniqueness that made us restrict the results of this paper to the grounded semantics. An extension of the work to other, potentially multiple extension semantics, is left for future work.

2.1. Value-Based Argumentation

In order to capture the specificity of argumentation about decisions, it is needed to take into account the values to which arguments appeal (e.g Bench-Capon, 2003; Bench-Capon, Doutre, & Dunne, 2007; Modgil, 2009). This approach is referred to as *value-based argumentation.*

Value-based argumentation assumes that an audience of a discussion can establish the relative strength of arguments on the basis of importance of values to which arguments appeal. Consequently, an attack on an argument appealing to a higher value than its attacker, can be disregarded by a relevant audience. As a result, some particular decision-makers can be persuaded by a given argumentation to a different extent than others.

Value-based argumentation frameworks are an extension of the abstract argumentation frameworks. In addition to the set of arguments and a binary attack relation, a set of values and a function mapping them to arguments are taken into account.

Definition 2.5 *A* value-based argumentation framework *(VAF) is a tuple $VAF = \langle A, \rightarrow , V, val \rangle$ where A is a set of arguments, $\rightarrow \subseteq A^2$ is an attack relation, V is a set of values and $val : A \rightarrow V$ is a function assigning values to arguments.*

The assignment of values to arguments provides a way of determining the strength of arguments from the perspective of a particular *audience*. This is done by the specification of an agent's preferences over values.

Definition 2.6 *Let $VAF = \langle A, \rightarrow, V, val \rangle$. An* audience *$P$ is a linear ordering over V. We denote that a value v_1 is strictly more preferable than a value v_2 for P as $v_1 >_P v_2$.*

Then, we say that an argument defeats another, if it attacks it and the value it is assigned is stronger or equal to the value of the attacked argument.

Definition 2.7 *Let $VAF = \langle A, \rightarrow, V, val \rangle$ be a VAF and P be an audience. Then, we say that an argument a* defeats *an argument b for P (we denote it as $a \rightarrow^P b$) iff $a \rightarrow b$ and it is not the case that $val(b) >_P val(a)$.*

Given this relation, we can consider a *defeat graph*, which is an argumentation framework consisting of the set of arguments and of the defeat relation for some ordering P.

Definition 2.8 *Let $VAF = \langle A, \rightarrow, V, val \rangle$ and P be an audience. The* defeat graph *of VAF based on P is an argumentation framework $VAF_P = \langle A, \rightarrow^P \rangle$.*

Let us illustrate the presented formalism on an example of a specific debate regarding making a practical decision.

Example 2.1 *(Airiau et al., 2016) Consider a debate regarding the possible ban of diesel cars, aimed at the reduction of air pollution in big cities. The following arguments are included in the discussion:*

- *A - Diesel cars should be banned.*
- *B - Artisans, who should be protected, cannot change their cars as it would be too expensive for them.*
- *C - We can subsidize electric cars for artisans.*
- *D - Electric cars, which could be a substitute for diesel, require too many new charging stations.*
- *E - We can build some charging stations.*
- *F - We cannot afford any additional costs.*
- *G - Health is more important than economy, so we should spend whatever is needed for fighting pollution.*

Further, it can be noticed that these arguments appeal to certain values. In particular, arguments A, G appeal to environmental responsibility (ER), B, C to social fairness (SF), F to economic viability (EV) and D, E - to infrastructure efficiency (IE).

These arguments are represented on the graph with a mapping of values depicted on Figure 1. For each argument, the first element of its description is its name, and the second one is the name of the value it appeals to [1] *.*

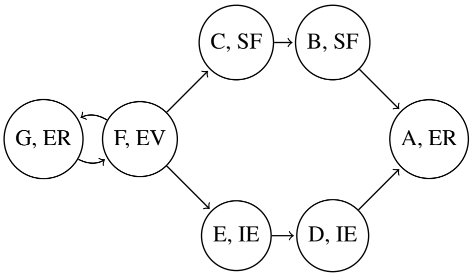
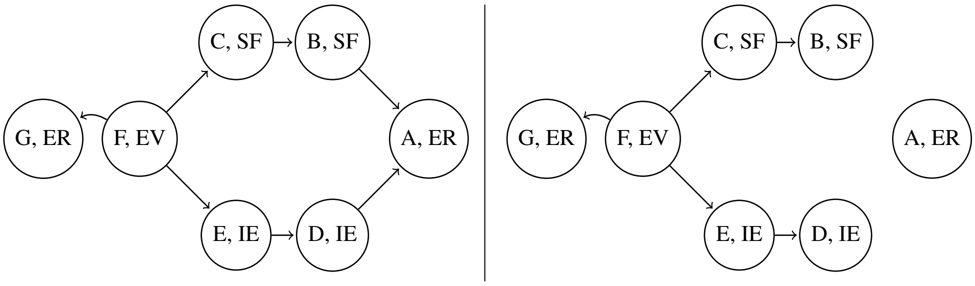


Figure 1. Value-based argumentation framework VAF of Example 2.1

[1] Notice that the name of the values arguments appeal to is left in the nodes of the defeat graph, for clarity of the representation, but that they are not part of the formal structure of the framework.

Let us now consider the structure of this discussion from the perspectives of two experts of a decision-making jury, that should decide on whether Diesel cars should be banned or not.

For Expert 1, economic viability is the most important. She ranks infrastructure efficiency lower, but higher than social fairness. Environmental responsibility is the least important for her. Then, from her point of view attacks in which the attacker appeals to a less important value than the attacked argument are disregarded. Taking her preferences into account, the structure presented in Figure 2 (a) is obtained, after the elimination of disregarded attacks. The grounded extension of this defeat graph is $\{F, B, D\}$.



(a) Expert 1's audience: EV > IE > SF > ER (b) Expert 2's audience: EV > ER > SF > IE

Figure 2. Defeat graphs based on (a) Expert 1's and (b) Expert 2's audiences

Let us now consider another expert of the jury, who believes that economic viability is the most important value. Expert 2 ranks environmental responsibility second, and social fairness third. Finally, she considers infrastructure efficiency as the least important. From her perspective, the structure of successful attacks is much different, as indicated in Figure 2 (b). The grounded extension of this defeat graph is $\{F, B, D, A\}$.

2.2. *Decisive Argument*

When aiming at reaching a decision, like in Example 2.1, an argument can be decisive, in the sense that if it states that a decision should be taken, and if it is an acceptable argument, then the decision should be taken.

This intuition can be captured by mapping information about a decisive support for a decision to a considered argument.

Definition 2.9 *Let* $AF = \langle A, \rightarrow \rangle$ *be an argumentation framework. We call* $DP = \langle AF, D \rangle$ *a* decision problem w.r.t. an AF, *where* $D \in A$ *is the* decisive argument. *Also, let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be a value-based argumentation framework. We call* $DP = \langle VAF, D \rangle$ *a* decision problem w.r.t. a VAF, *where* $D \in A$ *is the* decisive argument.

Example 2.2 *(Continuation of Example 2.1) The decision of the experts of the jury relies on whether the argument stating that the ban should be introduced, that is, on argument A, is accepted or not. This decisive argument should be considered for the VAF, for the defeat graphs of each of the experts, and for the collective defeat graph that will represent their common view.*

The notion of collective defeat graph will be defined in the next section.

3. Aggregation of Audiences in Value-Based Argumentation Frameworks

One of the intuitions behind the claim that the structure of argumentation can be viewed in distinctive ways is that agents might not agree on whether particular arguments are indeed in conflict with each other. Possible explanations for such a situation involve a scenario in which particular agents disagree on the relative strength of arguments. As it was argued earlier, it is plausible to assume that if an agent believes that some strong argument is attacked by a weak one, she might decide to disregard this attack. However, decisions about which arguments are stronger than another are at the discretion of individual assessors. Therefore, structures of *successful* attacks between arguments might vary among the group of agents.

One of the methods of aggregating views on the relative strength of arguments is to reach a collective view about the ordering over values that they appeal to. This approach makes use of *preference aggregation functions*, pioneered by Arrow (1951), which constitute one of major parts of social choice theory. In this way we might establish a collective ordering over values and consequently compute a collective defeat graph of the initial *VAF*. Then, evaluation of acceptance of a decisive argument can be performed. This approach has been proposed earlier in the context of value-based argumentation by Pu et al. (2013).

In order to provide the described procedure formally, preference aggregation functions will be used. This mechanism, widely studied in social choice theory, considers a profile of orderings over a set of items. Further, it provides a single, collective ordering.

Definition 3.1 (Preference Aggregation Function) *Let $V = \{v_1, \dots, v_n\}$ be a set of options, $\mathcal{N} = \{1, \dots, m\}$ be a set of agents, and $\mathcal{P}$ be the set of all linear orderings over V. A profile of orderings (denoted as **P**) is an element of $\mathcal{P}^m$. Then a* preference aggregation function *is a function $F : \mathcal{P}^m \to \mathcal{P}$. We denote the set of agents supporting $v_i > v_j$ in a profile of orderings **P** as $N_{\mathbf{P}}^{v_i > v_j}$.*

Then, given a $VAF = \langle A, \rightharpoonup, V, val \rangle$ and a profile **P** of linear orderings over V, the defeat graph $AF = \langle A, \rightharpoonup^{Borda(\mathbf{P})} \rangle$ can be considered as the collective defeat graph.

Definition 3.2 *Take a $VAF = \langle A, \rightharpoonup, V, val \rangle$, a profile **P** of preference orderings over V and a preference aggregation function F. Then, a* collective defeat graph *for **P** under F is the argumentation framework $AF = \langle A, \rightharpoonup^{F(\mathbf{P})} \rangle$.*

An example of such a preference aggregation function is the Borda rule. This rule is simple and easy to present which is why we use it for the illustration of the discussed mechanisms. Naturally, a variety of different rules can be used. To describe the Borda rule, let us also introduce a handy notation.

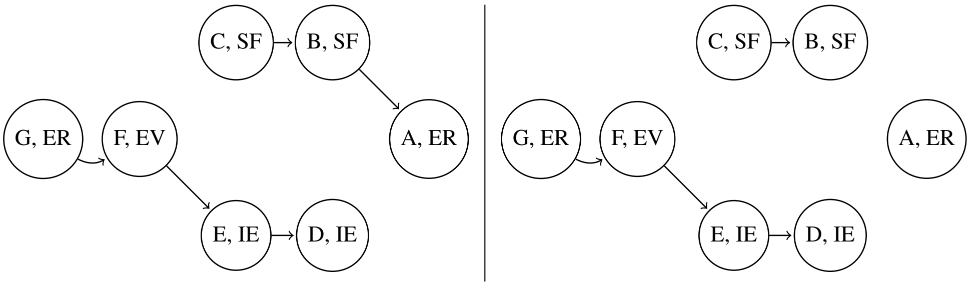
Notation 1 *Let P be a linear order over some set V. We denote by $top(P)$ the option $v \in V$ such that for any $v' \neq v$, $v >_P v'$. Further, we denote as $rank_P(v)$ the position of the option v in the ordering P. Formally, $rank_P(v) = |\{v' \in V | v' >_P v\}| + 1$.*

To compute the result of the Borda rule, for any element P_i of a profile of linear preference orderings **P** of length m over a set of options V, we assign to each option a number of points. A score of an option v_j given by an agent i, called $BordaScore_i(v_j)$ amounts to $|V| - rank_{P_i}(v_j)$. Then, an overall score of v_j, namely $BordaScore(v_j) = \Sigma_{i=1}^{n} BordaScore_i(v_j)$. Finally $Borda(\mathbf{P})$ is a preference ordering in which the rank of each option is determined by the number of gained points. To obtain a linear ordering as the output of this function additional tie-breaking rules are needed.

Example 3.1 *(Continuation of Example 2.2) Let us consider an additional expert, Expert 3. Let us present her audience, and let us recall the audiences of the other two experts. These three experts form a panel **P**.*

- *Expert 1: EV > IE > SF > ER*
- *Expert 2: EV > ER > SF > IE*
- *Expert 3: SF > ER > EV > IE*

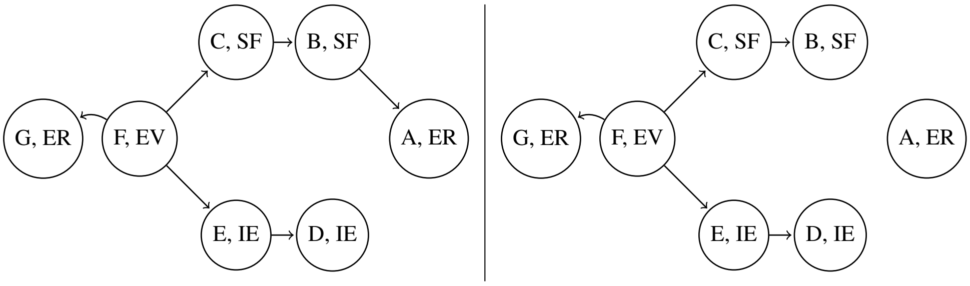
The defeat graph based on Expert 3's audience is depicted on Figure 3 (a).



(a) Sincere audience SF > ER > EV > IE (b) Insincere audience ER > SF > EV > IE

Figure 3. Defeat graphs based on Expert 3's audience

*Let us now calculate the result of the Borda rule for **P**. The scores are: ER: 4, EV: 7, IE: 2, SF: 5. So, Borda(**P**)= EV > SF > ER > IE. The defeat graph for this ordering is presented in Figure 4 (a); this is the collective defeat graph for the panel.*



(a) Before any manipulation (b) After manipulation by Expert 3

Figure 4. Collective defeat graph for the panel **P**, under the Borda rule

3.1. Manipulation in Aggregating Audiences

In the current work we are interested in the behavior of agents who are willing to misrepresent their sincere beliefs about the ranking of values in order to ensure that their desired decision is agreed upon. Such a behavior is called *manipulation*. An important assumption made for the sake of simplicity in the current work is that agents are aware of orderings over values in which other agents believe. Relaxing this assumption would be an interesting direction for further research.

To formalize this notion, let us define what agents' preferences over outcomes of aggregation are. In our setting agents are interested in ensuring that the collective preference ordering induces a defeat graph in which the decisive argument that supports their desired decision is accepted if and only if it is accepted in the defeat graph induced by agents' own ordering.

Following this intuition we say that given a decision problem and agents' ordering over values, an agent prefers some ordering to another if it treats the decisive argument consistently with the agent's intentions, while the other does not. Notice that these preferences are dichotomous. Given a decision problem $DP = \langle VAF, D \rangle$ and a preference ordering P_i corresponding to some agent i, we say that i is in favor of D if it is in the grounded extension of the defeat graph induced by P_i. If it is not, we say that i is against D.

Definition 3.3 *Let $DP = \langle VAF = \langle A, \rightarrow, V, val \rangle, D \rangle$ be a decision problem and* i *be an agent with a preference ordering P_{i}. If* i *is* in favor of *D, for any pair of preference orderings P_1, P_2, $P_1 >_{\mathrm{i}} P_2$ iff $D \in GRND_{\langle A, \rightarrow^{P_1} \rangle}$ while $D \notin GRND_{\langle A, \rightarrow^{P_2} \rangle}$. Also, if* i *is* against *D, $P_1 >_{\mathrm{i}} P_2$ iff $D \notin GRND_{\langle A, \rightarrow^{P_1} \rangle}$ while $D \in GRND_{\langle A, \rightarrow^{P_2} \rangle}$.*

Let us depict the problem of manipulation on an example.

Example 3.2 *(Continuation of Example 3.1) Expert 1 is against the decision, since argument A is not in the grounded extension of the defeat graph based on her audience. Expert 2 is in favor of the decision. Regarding Expert 3, the defeat graph based on her audience (see Figure 3 (a)) has $\{G, C, E, A\}$ as grounded extension: Expert 3 also is in favor of the decision.*

The decisive argument A is not in $\{F, B, D\}$, the grounded extension of the collective defeat graph of Figure 4 (a).

Expert 3 may however decide to lie about her preference over values and submit an ordering: ER > SF > EV > IE. Now, the Borda scores would amount to: ER: 5, EV: 7, IE: 2, SF: 4. The modified result of the Borda rule is: EV > ER > SF > IE. The corresponding defeat graph is depicted on Figure 4 (b). The grounded extension of the defeat graph based on this ordering is $\{F, B, D, A\}$; the decisive argument belongs to this extension.

So, by misrepresenting her sincere beliefs, Expert 3 can ensure that the argument she is in favor of, is accepted in the collective defeat graph. Expert 3 can thus manipulate.

A preference aggregation rule is said to be *strategy-proof*, when it is never possible for a single agent to manipulate.

Definition 3.4 (Strategy-proofness with respect to argumentation) *A preference aggregation rule F is* strategy-proof with respect to argumentation *iff for any profile of preference orderings **P** any agent i and any preference ordering P_i^*, it is not the case that $F(P_i^*, \boldsymbol{P}_{-P_i}) >_i F(\boldsymbol{P})$, where $(P_i^*, \boldsymbol{P}_{-P_i})$ is the profile of orderings identical to **P** except of the replacement of P_i with P_i^*.*

Let us also phrase manipulation as a computational problem.

VAF-MANIPULATION(F)

Instance: $DP = \langle VAF, D \rangle$, a profile of preference orderings $\mathbf{P}$, agent i.
Question: Is there a preference ordering P_i^* such that $F(P_i^*, \mathbf{P}_{-P_i}) >_i F(\mathbf{P})$?

The results concerning manipulation in the argumentation setting will be based on the facts about manipulation in *voting mechanisms*. They will be introduced in the subsequent section.

3.2. Voting Mechanisms

Aggregating preference orderings is strictly connected with engineering voting rules. There, a group of voters elects an option out of a set of candidates. Mechanisms of this kind aim at ensuring that the winner of the elections represents agents' preferences accurately. Technically, a voting rule is a function $F : \mathcal{P}^m \to O$, where $\mathcal{P}$ is the set of all preference orderings over the set of options O and m is the number of voters. Notice that we have imposed that a rule always selects a single option. This property is known as the *resoluteness* condition.

Voting rules can be envisaged as preference aggregation rules. Then, the winner of elections is the top option of the collective preference ordering.

If this is the case, preferences of particular voters can be clearly defined. Each of them wants to make sure that the winner of the election is as good as possible from the perspective of their ranking.

Definition 3.5 (Strategic Voting Preferences) *Let an agent i submit some ordering P_i over some set of options V. Then, for any pair of preference orderings P_1, P_2 over V, $P_1 >_i^V P_2$ iff $rank_{P_i}(top(P_1)) > rank_{P_i}(top(P_2))$.*
Then, we can ask if an agent can make herself better off with respect to strategic voting preferences by submitting an insincere preference ordering. If for some function F it is never the case, we say that F is strategy-proof with respect to voting preferences.

Definition 3.6 (Strategy-proofness in voting) *A preference aggregation rule F is* strategy-proof in voting *iff for any profile of preference orderings $\boldsymbol{P}$, any agent i and any preference ordering P_i^*, it is not the case that $F(P_i^*, \boldsymbol{P}_{-P_i}) >_i^V F(\boldsymbol{P})$.*

The Gibbard-Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975) states that any rule which is strategy-proof with respect to voting preferences is also dictatorial with respect to strategic voting. Here, a rule is said to be dictatorial if there is an agent whose *most preferred* option is always selected. Conditions of the theorem involve nonimposition, which means that any option is elected by some preference ordering. Also, the conditions include resoluteness.

Theorem 1 (Gibbard - Satterthwaite) *Any resolute, nonimposed, and strategy-proof voting rule for three or more alternatives must be a dictatorship.*

For instance, Borda rule cannot be strategy-proof as it is not a dictatorship.

3.3. Application of Strategic Voting

We will show that if it is possible for an agent to manipulate a preference aggregation rule with respect to voting, she might also manipulate with respect to argumentation.

Proposition 1 *Any preference aggregation rule F which is manipulable with respect to voting preferences is also manipulable with respect to argumentation.*

PROOF. *Consider any preference aggregation rule F which is manipulable with respect to strategic voting. This means that there is a set of voters $N = \{1,\dots,n\}$, a set of options $V = \{v_1,\dots,v_m\}$ and a profile of preference orderings submitted by voters $\boldsymbol{P} = \langle P_1,\dots,P_n\rangle$ such that for some voter i, there is some preference ordering P_i^* over V such that $rank_i(top(F(\boldsymbol{P}))) < rank_i(top(F(\boldsymbol{P}^*)))$, where $\mathbf{P}^*$ is $\mathbf{P}$ with P_i replaced by P_i^*. Take such a profile. We will construct a decision problem $DP = \langle VAF, C\rangle$ which is manipulable by the successful manipulator with respect to strategic voter.*

Let us take a set of values V and the set of arguments $A = \{a_1,\dots,a_m\}$ (one per element of V). Further, take the valuation map val such that for any $a_i \in A$, $val(i) = v_i$. For simplicity let us say that $P_i = v_1 \succ_i v_2 \succ_i \cdots \succ_i v_m$. Now, let a_1 be the decisive argument. Then, let v_j correspond to top(F($\mathbf{P}$)). Construct the attack relation so that $a_j \rightarrow a_1$. Also, for any v_b such that $rank_i(v_b) > rank_i(j)$, let $a_b \rightarrow a_j$. No other attacks are considered.

Then firstly notice that for the agent i the argument a_1 should be accepted, as it is in the grounded extension of the defeat graph based on i's preference ordering. However, it is not included in the grounded extension of the defeat graph based on $F(\boldsymbol{P})$, as all attacks on a_j are eliminated because v_j is the top value. However, we know that i can submit an ordering P^* such that some $v_b \succ_i v_j$ becomes the top option. Then, clearly one of the attackers of v_j, which is the only attacker of the decisive argument is the top option, so the attack is preserved. Therefore, v_i is accepted in the new defeat graph. So i manipulated successfully. ■

This result is followed by an unfortunate conclusion. Namely, it turns out that for any preference aggregation rule F based on strict preferences, if F is not dictatorial with respect to strategic voting, it is manipulable in the current setting.

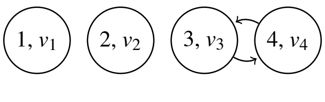
To justify this claim it is sufficient to take any preference aggregation rule F based on strict preferences and suppose that it is not dictatorial with respect to strategic voting. Then, by the Gibbard-Satterthwaite theorem we know that it is manipulable with respect to strategic voting. But then it follows that it is also manipulable in the argumentation setting.

This means that aggregating *VAF*s by preference aggregation is at least as vulnerable to strategic behavior as voting mechanisms. In fact, we can find cases in which rules strategy-proof with respect to strategic voting are manipulable within the argumentation setting.

Observation 1 *It is not true that if a rule F is strategy-proof in voting, it is strategy-proof with respect to argumentation.*

Example 3.3 *Consider the following preference aggregation rule F: For any profile of preference orderings $\boldsymbol{P}$ distinguish a dictator d. Then, $top(F(\boldsymbol{P})) = top(\boldsymbol{P}_d)$. To determine the rest of the collective preference ordering, eliminate the value $top(P_d)$ from the profile of orderings. Apply the Borda rule to the remainder of the profile. This rule is strategy-proof with respect to strategic voting, as the top option is known from the start.*

Now consider the following decision problem DP, where argument 3 is the decisive argument:



Now take a profile $\boldsymbol{P}$, where agent d is the dictator:
$\boldsymbol{d}: v_2 \succ v_1 \succ v_4 \succ v_3$, $\boldsymbol{m}: v_3 \succ v_4 \succ v_1 \succ v_2$, $\boldsymbol{o}: v_4 \succ v_3 \succ v_2 \succ v_1$

Let us now notice that the score of v_1 is 3, v_3 receives 5 points, and v_4 gets 6 . The score of v_2 does not matter as it is dictator's top option. So, we get $F(\boldsymbol{P}) = v_2 \succ v_4 \succ v_3 \succ v_1$. It is easy to see that under this ordering argument 3 is not in the grounded extension, as the attack $4 \to 3$ is preserved, while $3 \to 4$ is not. This leaves agent m dissatisfied, as in the defeat graph based on her preferences argument 3 is clearly accepted.

Consider, however, the profile:
$\boldsymbol{d}: v_2 \succ v_1 \succ v_4 \succ v_3$, $\boldsymbol{m}^*: v_3 \succ v_2 \succ v_1 \succ v_4$, $\boldsymbol{o}: v_4 \succ v_3 \succ v_2 \succ v_1$

After this change the score of v_1 is 3, v_3 receives 5 points, and v_4 gets 6. Therefore, the collective preference ordering is $v_2 \succ v_4 \succ v_3 \succ v_1$. Under this ordering 3 is accepted in the collective structure. So, m manipulated successfully.

This means that the current setting is strictly less immune to strategic behavior than strategic voting. However, we can show that for a large class of rules, the manipulation problem is difficult to compute. Thanks to this observation we can claim that the proposed mechanism can eliminate manipulation in practical applications. As we will show, if a rule is NP-hard with respect to the problem of strategic voting, it is also NP-hard with respect to our setting.

Proposition 2 *For any preference aggregation rule F for which the strategic voting problem is NP-hard, so is the VAF-manipulation problem with respect to F.*

PROOF. *Take any preference aggregation rule F for which the strategic voting problem is NP-hard. Let us show the way to reduce this problem to manipulation in the current setting. Take a profile of preference orderings $\boldsymbol{P}$ and an agent i with a preference ordering P_i. Let us construct a decision problem in which i can manipulate if and only if she can manipulate with respect to strategic voting. Take a VAF in which we have an argument corresponding to any ranked option. Also, map each of the options as values of corresponding arguments. Further, let i's favourite option correspond to the decisive argument - a_i. Now, let the argument a_j, corresponding to $top(F(\boldsymbol{P}))$ attack a_i iff $a_i \neq a_j$. Also, let any argument a_b such that $val(a_b) \succ_{P_i} val(a_i)$ attack a_j. Clearly, i is in favor of a_i. We need to show that i can manipulate with respect to argumentation setting iff she can manipulate with respect to strategic voting. If i can manipulate with respect to argumentation setting, then there is a preference ordering P_i^* such that $rank_i(top(F(P_i^*, \boldsymbol{P}_{-P_i}))) > rank_i(top(F(\boldsymbol{P})))$. Otherwise, a_i would not be in the grounded extension of the defeat graph induced by $F(P_i^*, \boldsymbol{P}_{-P_i})$. But then, i can manipulate with respect to strategic voting. But also, if there is a preference ordering P_i^* such that $rank_i(top(F(P_i^*, \boldsymbol{P}_{-P_i}))) > rank_i(top(F\boldsymbol{P}))$, then D becomes in the grounded extension of the defeat graph induced by $F(P_i^*, \boldsymbol{P}_{-P_i})$. So, i can manipulate with respect to VAF-manipulation.* ■

4. Conclusions

In this work we studied applications of social choice mechanisms to aggregating views on preferences over values. Following Pu et al. (2013), we used preference aggregation functions to determine a collective preference ordering over values. Further, we have studied strategic behavior within the proposed models for collective decision-making. We used results concerning strategic voting to establish conditions for manipulability in preference aggregation. Following this connection we also obtained results concerning the complexity of manipulation problem in the preference aggregation approach.

We have shown that strictly more rules are manipulable with respect to the studied decision-making setting than with respect to strategic voting. This means, that any rule strategy-proof with respect to aggregating audiences in *VAF*s is dictatorial with respect to strategic voting. It is worth noting that this is not necessarily a very problematic result. A strategy-proof preference aggregation function which is only dictatorial with respect to one value can still be fair with respect to a large part of the preference ordering.

Our study leaves room for further work. It would be interesting to study the complexity of manipulating preference aggregation by submitting a preference ordering which is minimally different from the agents' sincere hierarchy of values. Also, it would be beneficial to investigate the complexity of manipulation for semantics other than grounded. Another interesting avenue of research would be to investigate structural properties of *VAF*s eliminating the possibility of strategic behavior.

References

Airiau, S., Bonzon, E., Endriss, U., Maudet, N., & Rossit, J. (2016). Rationalisation of Profiles of Abstract Argumentation Frameworks. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multi-Agent Systems (AAMAS)* (pp. 350–357).

Amgoud, L., & Cayrol, C. (2002). A Reasoning Model Based on the Production of Acceptable Arguments. *Ann. Math. Artif. Intell.*, *34*(1-3), 197–215.

Arrow, K. J. (1951). Social Choice and Individual Values.

Bench-Capon, T. (2003). Persuasion in Practical Argument Using Value-Based Argumentation Frameworks. *Journal of Logic and Computation*, *13*(3), 429–448.

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

Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, *77*(2), 321–357.

Gibbard, A. (1973). Manipulation of Voting Schemes: a General Result. *Econometrica: Journal of the Econometric Society*, 587–601.

Modgil, S. (2009). Reasoning About Preferences in Argumentation Frameworks. *Artificial Intelligence*, *173*(9-10), 901–934.

Pu, F., Luo, J., Zhang, Y., & Luo, G. (2013). Social Welfare Semantics for Value-Based Argumentation Framework. In *Proceedings of International Conference on Knowledge, Science, Engineering and Management* (pp. 76–88).

Rahwan, I., & Larson, K. (2008). Mechanism Design for Abstract Argumentation. In *Proceedings of the 2008 International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 1031–1038).

Satterthwaite, M. A. (1975). Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, *10*(2), 187–217.