

# Collective Argument Evaluation as Judgement Aggregation

Iyad Rahwan

<sup>1</sup>Masdar Institute of Science & Technology, UAE

<sup>2</sup>(Visiting Scholar) Massachusetts Institute of Technology, Cambridge MA, USA

<sup>3</sup>(Fellow) University of Edinburgh, UK\*

Fernando Tohmé

Artificial Intelligence Research and Development Lab (LIDIA), Universidad Nacional del Sur, Bahía Blanca, CONICET, Argentina

## ABSTRACT

A conflicting knowledge base can be seen abstractly as a set of arguments and a binary relation characterising conflict among them. There may be multiple plausible ways to evaluate conflicting arguments. In this paper, we ask: *given a set of agents, each with a legitimate subjective evaluation of a set of arguments, how can they reach a collective evaluation of those arguments?* After formally defining this problem, we extensively analyse an argument-wise plurality voting rule, showing that it suffers a fundamental limitation. Then we demonstrate, through a general impossibility result, that this limitation is more fundamentally rooted. Finally, we show how this impossibility result can be circumvented by additional domain restrictions.

## Categories and Subject Descriptors

I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—*multiagent systems, coherence and coordination*

## General Terms

Theory, Economics

## Keywords

Argumentation, Social Choice, Judgement Aggregation

## 1. INTRODUCTION

Argumentation has recently become one of the key approaches to automating and analysing reasoning in the presence of conflicting information. A key milestone in the development of argumentation in AI has been Dung's landmark framework [8]. Arguments are viewed as abstract entities, with a binary defeat relation among them (resulting in a so-called *argument graph*).

Often, there are multiple reasonable ways in which an agent may evaluate a given argument graph. Each possible evaluation corresponds to a so-called *extension* [8] or

\*This work commenced when Iyad Rahwan was with the British University in Dubai.

**Cite as:** Collective Argument Evaluation as Judgement Aggregation, I. Rahwan and F. Tohmé, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

Copyright © 2010, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

*labelling* [5]. We ask: *Given an argument structure and a set of agents, each with a legitimate subjective evaluation of the given arguments, how can the agents reach a collective compromise on the evaluation of those arguments?*

We formally define the problem of aggregating multiple evaluations of arguments, in the spirit of preference aggregation [1] and judgement aggregation [10, 11]. We define a specific aggregation operator (argument-wise plurality voting) and analyse some of its key properties. We then present an impossibility result on the existence of good aggregation operators (in particular, satisfying collective rationality). Then, we show one way in which the impossibility result can be avoided. In particular, we provide a full characterisation of the space of individual judgements that guarantees collective rationality using argument-wise plurality voting.

The paper makes three key contributions to the state-of-the-art in computational models of argument. Firstly, the paper defines and analyses the argument-wise plurality voting mechanism for collective argument evaluation.

Our second contribution is a general impossibility result, showing that there is no aggregation operator that can satisfy a few simple requirements (common in social choice theory) for arbitrary argument graphs. This result not only helps us avoid the fruitless pursuit of such operator, but also because it motivates the need for specialised aggregation operators that work under more restrictive conditions.

This leads to the third contribution. By showing how the impossibility result can be avoided by restricting the space of possible individual judgements, we provide guidance on circumventing the practical implications of the problem.

## 2. PRELIMINARIES

We briefly outline key elements of abstract argumentation frameworks [8], assuming finite sets of arguments.

**DEFINITION 1.** *An argumentation framework is a pair  $AF = \langle \mathcal{A}, \rightarrow \rangle$  where  $\mathcal{A}$  is a finite set of arguments and  $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$  is a defeat relation. We say that an argument  $\alpha$  defeats an argument  $\beta$  if  $(\alpha, \beta) \in \rightarrow$  (also written  $\alpha \rightarrow \beta$ ).*

An argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterise defeat among arguments. An example argument graph is shown in Figure 1. Argument  $\alpha_1$  has two defeaters (*i.e.* counter-arguments)  $\alpha_2$  and  $\alpha_4$ , which are themselves defeated by arguments  $\alpha_3$  and  $\alpha_5$  respectively. Let  $S^+ = \{\beta \in \mathcal{A} \mid \alpha \rightarrow \beta \text{ for some } \alpha \in S\}$ . Also let  $\alpha^- = \{\beta \in \mathcal{A} \mid \beta \rightarrow \alpha\}$ . We first characterise the fundamental notions of conflict-free and defence.

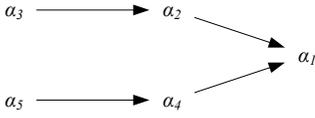


Figure 1: A simple argument graph

DEFINITION 2. Let  $\langle \mathcal{A}, \rightarrow \rangle$  be an argumentation framework, let  $S \subseteq \mathcal{A}$  and  $\alpha \in \mathcal{A}$ .

- $S$  is conflict-free iff  $S \cap S^+ = \emptyset$ .
- $S$  defends argument  $\alpha$  iff  $\alpha^- \subseteq S^+$ . Equivalently, we say that argument  $\alpha$  is acceptable with respect to  $S$ .

Intuitively, a set of arguments is *conflict free* if no argument in that set defeats another. A set of arguments *defends* a given argument if it defeats all its defeaters. In Figure 1, for example,  $\{\alpha_3, \alpha_5\}$  defends  $\alpha_1$ . We now look at some ways to characterise the *collective acceptability* of a set of arguments.

DEFINITION 3 (CHARACTERISTIC FUNCTION). Let  $AF = \langle \mathcal{A}, \rightarrow \rangle$  be an argumentation framework. The characteristic function of  $AF$  is  $\mathcal{F}_{AF}: 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$  such that, given  $S \subseteq \mathcal{A}$ , we have  $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} \mid S \text{ defends } \alpha\}$ .

When there is no ambiguity about the argumentation framework in question, we will use  $\mathcal{F}$  instead of  $\mathcal{F}_{AF}$ .

DEFINITION 4. Let  $S$  be a conflict-free set of arguments in framework  $\langle \mathcal{A}, \rightarrow \rangle$ .

- $S$  is admissible iff it is conflict-free and defends every element in  $S$  (i.e. iff  $S \subseteq \mathcal{F}(S)$ ).
- $S$  is a complete extension iff  $S = \mathcal{F}(S)$ .

Intuitively, a set of arguments is *admissible* if it is a conflict-free set that defends itself against any defeater – in other words, if it is a conflict free set in which each argument is acceptable with respect to the set itself.

An admissible set  $S$  is a *complete extension* if and only if all arguments defended by  $S$  are also in  $S$  (that is, if  $S$  is a fixed point of the operator  $\mathcal{F}$ ). There may be more than one complete extension, each corresponding to a particular consistent and self-defending viewpoint.

EXAMPLE 1. In Figure 1, the sets  $\emptyset$ ,  $\{\alpha_3\}$ ,  $\{\alpha_5\}$ , and  $\{\alpha_3, \alpha_5\}$  are all admissible simply because they do not have any defeaters. The set  $\{\alpha_1, \alpha_3, \alpha_5\}$  is also admissible since it defends itself against both defeaters  $\alpha_2$  and  $\alpha_4$ . The admissible set  $\{\alpha_1, \alpha_3, \alpha_5\}$  is the only complete extension, since  $\mathcal{F}(\{\alpha_1, \alpha_3, \alpha_5\}) = \{\alpha_1, \alpha_3, \alpha_5\}$ .

There are various approaches to differentiate between different complete extensions (e.g. by defining grounded, preferred, stable extensions and so on [8]). In this paper, we will take a liberal approach and consider any complete extension as a reasonable point of view for an agent, satisfying the minimal criteria of consistency and self-defence.

Crucial to our subsequent analysis is the notion of *argument labelling* [5]. It specifies which arguments are accepted (labelled **in**), which ones are rejected (labelled **out**), and which ones whose acceptance or rejection could not be decided (labelled **undec**). Labellings must satisfy two conditions: (i) an argument is **in** if and only if all of its defeaters are **out**; (ii) an argument is **out** if and only if at least one of its defeaters is **in**.

DEFINITION 5 (ARGUMENT LABELLING). Let  $AF = \langle \mathcal{A}, \rightarrow \rangle$  be an argumentation framework. An argument labelling is a total function  $L: \mathcal{A} \rightarrow \{\text{in}, \text{out}, \text{undec}\}$  such that:

- $\forall \alpha \in \mathcal{A} : (L(\alpha) = \text{out} \equiv \exists \beta \in \mathcal{A} \text{ such that } (\beta \rightarrow \alpha \text{ and } L(\beta) = \text{in}))$ ; and
- $\forall \alpha \in \mathcal{A} : (L(\alpha) = \text{in} \equiv \forall \beta \in \mathcal{A} : (\text{if } \beta \rightarrow \alpha \text{ then } L(\beta) = \text{out}))$

If none of the two conditions is satisfied, then  $L(\alpha) = \text{undec}$  (since  $L$  is a total function).

Caminada [5] showed a one-to-one correspondence between possible labellings and the set of all complete extensions.

### 3. MOTIVATION AND SCOPE

In this section, we give a simple example and use it to motivate the paper and highlight the scope of its contributions. Consider the following simple example.

EXAMPLE 2 (A MURDER CASE). A murder case is under investigation. To start with, there is an argument that the suspect should be presumed innocent ( $\alpha_3$ ). However, there is evidence that he may have been at the crime scene at the time ( $\alpha_2$ ), which would counter the initial presumption of innocence. There is also, however, evidence that the suspect was attending a party that day ( $\alpha_1$ ). Clearly,  $\alpha_1$  and  $\alpha_2$  are mutually defeating arguments since the suspect can only be in one place at any given time. This problem can be modelled as an argumentation framework  $AF = \langle \{\alpha_1, \alpha_2, \alpha_3\}, \rightarrow \rangle$  with  $\rightarrow = \{(\alpha_1, \alpha_2), (\alpha_2, \alpha_1), (\alpha_2, \alpha_3)\}$ . Possible labellings are:

- $L(\alpha_1) = \text{in}, L(\alpha_2) = \text{out}, L(\alpha_3) = \text{in}$ .
- $L'(\alpha_1) = \text{out}, L'(\alpha_2) = \text{in}, L'(\alpha_3) = \text{out}$ .
- $L''(\alpha_1) = \text{undec}, L''(\alpha_2) = \text{undec}, L''(\alpha_3) = \text{undec}$ .

The graph and possible labellings are depicted in Figure 2.

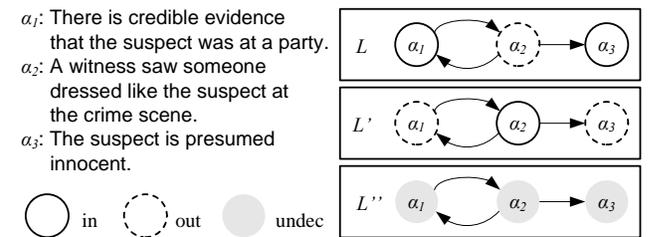


Figure 2: Graph with three possible labellings

Example 2 highlights a situation in which multiple points of view can be taken, depending on whether one decides to accept the argument that the suspect was at the party or the crime scene. Consider the following example.

EXAMPLE 3 (THREE DETECTIVES). A team of three detectives, named 1, 2, and 3, have been assigned to the murder case described in Example 2. Each detective's judgement can only correspond to a legal labelling (otherwise, his/her judgement is not admissible and can be discarded). Suppose that each detective's judgement is such that  $L_1 = L$ ,  $L_2 = L'$  and  $L_3 = L'$ . That is, detectives 2 and 3 agree but differ with detective 1. These labellings are depicted in the labelled graph of Figure 3. The detectives must decide which (aggregated) argument labelling best reflects their collective judgement.

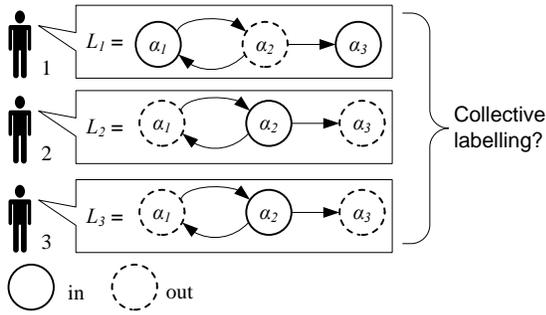


Figure 3: Detectives with different judgements

Example 3 highlights an aggregation problem, similar to the problems of preference aggregation [1] and judgement aggregation [10]. It is perhaps obvious in this particular example that  $\alpha_3$  must be rejected (and thus the defendant be considered guilty), since most detectives seem to think so. For the same reason,  $\alpha_1$  must be rejected and  $\alpha_2$  must be accepted. Thus, labelling  $L'$  (see Example 2) wins by majority. As we shall see in our analysis below, things are not that simple, and counter-intuitive situations may arise. To summarise, the question is as follows: *Given a set of agents, each with a specific subjective labelling of a given set of conflicting arguments, how can agents reach a collective decision on how to evaluate those arguments?*

Below, we will explore the above question deeply. We introduce the argument-wise plurality voting rule and study its key properties. We show that while argument-wise plurality voting satisfies many desirable properties (*e.g.* anonymity, strategy-proofness *etc.*), it can produce counter-intuitive results. We then generalise this observation by presenting a general impossibility result on the existence of collectively rational aggregation operators for argument labelling. We then fully characterise restrictions on the space of individual judgements under which the argument-wise plurality voting avoids the impossibility result.

## 4. AGGREGATION OF LABELLINGS

The problem we face is that of *judgement aggregation* [10] in the context of argumentation frameworks. In particular, taking as an input a set of *individual* judgements as to how each argument in  $AF$  must be labelled, we need to come up with a *collective* judgement. If each agent  $i = 1, \dots, n$  has a labelling  $L_i$ , we need to find an *aggregation operator*, which we define as a partial function<sup>1</sup>  $F : \mathbf{L}(AF)^n \rightarrow \{\text{in}, \text{out}, \text{undec}\}^{\mathcal{A}}$ , where  $\mathbf{L}(AF)$  is the class of labellings of  $AF$ . This means that for each  $\alpha \in \mathcal{A}$ ,  $F(L_1, \dots, L_n)[\alpha]$  is the label assigned to  $\alpha$  (if  $F$  is defined for  $\alpha$ ).

Aggregation involves comparing and assessing different points of view. There are, of course, many ways of doing this, as extensively discussed in the literature of Social Choice Theory [9]. In this literature, a consensus on some normative ideals has been reached, identifying what a ‘fair’ way of adding up preferences should be. So for instance, if everybody agrees, the outcome must reflect that agreement; no single agent can impose her view on the aggregate; the

<sup>1</sup>We state that the function is partial to allow for cases in which collective judgement may be undefined (*e.g.* when there is a tie in voting).

aggregation should be performed in the same way in each possible case, *etc.* These informal requirements can be formally stated as properties that  $F$  should satisfy [10, 7]:

Let  $AF = \langle \mathcal{A}, \rightarrow \rangle$  be an argumentation framework, and suppose we have  $n$  agents.

1. *Universal Domain*: Every possible profile of labellings  $(L_1, \dots, L_n)$  is in the domain of  $F$ .
2. *Unanimity*: If  $L_i = L$  for  $i = 1, \dots, n$ , then  $F(L_1, \dots, L_n) = L$ .
3. *Anonymity*: given any permutation  $p : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ ,  $F(L_1, \dots, L_i, \dots, L_n) = F(L_{p(1)}, \dots, L_{p(i)}, \dots, L_{p(n)})$ .
4. *Independence*: for any  $\alpha \in \mathcal{A}$ , and any two profiles  $(L_1, \dots, L_n)$  and  $(L'_1, \dots, L'_n)$ , if  $\forall i$  we have:  $L_i(\alpha) = l_\alpha$  iff  $L'_i(\alpha) = l_\alpha$ , then  $F(L_1, \dots, L_n)[\alpha] = l_\alpha$  iff  $F(L'_1, \dots, L'_n)[\alpha] = l_\alpha$ .
5. *Systematicity*: for any  $\alpha, \beta \in \mathcal{A}$  and any two profiles  $(L_1, \dots, L_n)$  and  $(L'_1, \dots, L'_n)$ , if  $\forall i, L_i(\alpha) = L'_i(\beta)$  then  $F(L_1, \dots, L_n)[\alpha] = F(L'_1, \dots, L'_n)[\beta]$ .
6. *Monotonicity*: For any  $\alpha \in \mathcal{A}$ ,  $l_\alpha \in \{\text{in}, \text{out}, \text{undec}\}$  is such that given two profiles  $(L_1, \dots, L_i, \dots, L_n)$  and  $(L_1, \dots, L'_i, \dots, L_n)$  (differing only in  $i$ 's labelling), if  $L_i(\alpha) \neq l_\alpha$  while  $L'_i(\alpha) = l_\alpha$ ,  $F(L_1, \dots, L_n)[\alpha] = l_\alpha$  implies that  $F(L_1, \dots, L'_i, \dots, L_n)[\alpha] = l_\alpha$ .
7. *Non-dictatorship*: there is no  $i$  such that for any profile  $(L_1, \dots, L_i, \dots, L_n)$ ,  $F(L_1, \dots, L_i, \dots, L_n) = L_i$ .
8. *Collective Rationality*:  $F(L_1, \dots, L_n)$ , is a labelling.

In words, *universal domain* requires that  $F$  admits any logically possible profile of agent judgements. *Unanimity* requires that if all agents submit the same labelling, this labelling must be the collective one. *Anonymity* means that all agents should have equal weight in the aggregation. *Independence* means that collective judgement on each argument should only depend on individual judgements about that particular argument. *Systematicity* combines independence with neutrality across arguments. *Monotonicity* states that if an agent switches its judgement on an argument in favour of the collective judgement, then the collective judgement remains the same. *Non-dictatorship* means that no single agent should always determine the collective judgement. *Collective rationality* means that the aggregation is always a legitimate labelling.

Notice that these conditions are not independent since, for instance, *Systematicity* implies *Independence* (just by choosing  $\alpha \equiv \beta$ ), but they reflect many properties that researchers consider a ‘good’ aggregation operator should have.

## 5. ARGUMENT-WISE PLURALITY VOTING

An obvious candidate aggregation operator to check out is the *plurality* voting operator  $M$ . In this section, we analyse a number of key properties of this operator. Intuitively, for each argument, it selects the label that appears most frequently in the individual labellings.

DEFINITION 6 (ARGUMENT-WISE PLURALITY). Let  $AF = \langle \mathcal{A}, \rightarrow \rangle$  be an argumentation framework. Given  $\alpha \in \mathcal{A}$ , then  $M(L_1, \dots, L_n)[\alpha] = l_\alpha \in \{\text{in}, \text{out}, \text{undec}\}$  iff

$$|\{i : L_i(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{i : L_i(\alpha) = l'_\alpha\}|$$

Otherwise,  $M(L_1, \dots, L_n)[\alpha] = \emptyset$ .

EXAMPLE 4 (THREE DETECTIVES (CONT.)). Continuing on Example 3, applying argument-wise plurality:

- $M(L_1, L_2, L_3)[\alpha_1] = \text{out}$
- $M(L_1, L_2, L_3)[\alpha_2] = \text{in}$
- $M(L_1, L_2, L_3)[\alpha_3] = \text{out}$

Note that even in the case of ties,  $M$  is well-defined since  $\emptyset$  is a member of every set. However, when  $M(L_1, \dots, L_n)[\alpha] = \emptyset$  for some  $\alpha \in \mathcal{A}$ , then the output of  $M$  is obviously not a legal labelling (*i.e.* Collective Rationality will be violated).

## 5.1 Strategic Manipulation

First, we ask whether the plurality aggregation rule is *strategy-proof*. Before such analysis can be done, it is important to define what might *motivate* agents to behave strategically, *i.e.* agent's preferences over labellings.

We define agents' preferences with respect to restricted sets of arguments in order to model situations where agents have potentially different *domains of knowledge*. As a motivating example, consider a court case where a medical expert is called as an expert witness. This expert can put forward arguments related to medical forensics, but would be unable to comment on legal issues. Similarly, an agent's arguments can be limited by their *position to know*. For example, a friend may be in a position to comment on someone's character, while a stranger's comments would not be of interest.

Let  $\theta_i \in \Theta_i$  denote the *type* of agent  $i \in I$  which is drawn from some set of possible types  $\Theta_i$ . The type represents the private information and preferences of the agent. More precisely,  $\theta_i$  determines agent  $i$ 's preferences are over *outcomes*  $L \in \mathcal{L}$ . By  $L_1 \succeq_i L_2$  we denote that agent  $i$  *weakly prefers* (or simply *prefers*) outcome  $L_1$  to  $L_2$ . We say that agent  $i$  *strictly prefers* outcome  $L_1$  to  $L_2$ , written  $L_1 \succ_i L_2$ , if and only if  $L_1 \succeq_i L_2$  but not  $L_2 \succeq_i L_1$ . Finally, we say that agent  $i$  is *indifferent* between outcomes  $L_1$  and  $L_2$ , written  $L_1 \sim_i L_2$ , if and only if both  $L_1 \succeq_i L_2$  and  $L_2 \succeq_i L_1$ .

Here, we consider *focal-set-oriented* agents. These agents have a core set of arguments which they care about, and their only interest is in their exact judgement on those arguments being adopted by the collective.

DEFINITION 7 (FOCAL-SET-ORIENTED). An agent  $i$  with labelling  $L_i$  is focal-set-oriented if there is a set of arguments  $\bar{A}^i \subseteq \mathcal{A}$ , called  $i$ 's focal-set, such that for any labelling  $L$ :

1.  $L_i \sim_i L$  iff  $\forall \alpha \in \bar{A}^i, L_i(\alpha) = L(\alpha)$ ;
2.  $L_i \succeq L$  otherwise.

Focal-set-orientation defines a very general class of agent preferences. An example of a focal-set-oriented agent is a *resolute agent*, that is only satisfied if the aggregated labelling exactly matches its own labelling. At the other extreme is an *agent with a focal argument*, which only cares

about the final status of a single argument. In this case, the agent's focal-set includes a single argument only.

Strategy-proofness (also known as *dominant strategy incentive compatibility*) is an important property in analysing agents' strategic incentives [12, page 871].<sup>2</sup> In our context, it asks whether any agent has incentive to misreport its labelling, given any possible reported labellings by other agents. Let  $L_{-i} = \{L_1, \dots, L_n\} \setminus L_i$  denote the set of labellings of agents other than agent  $i$ .

DEFINITION 8 (STRATEGY-PROOF). Let  $i \in I$  be an arbitrary agent with a labelling  $L_i$ .  $F$  is a strategy-proof aggregation operator iff  $\forall L_{-i}, \forall L_i^* \neq L_i, F(L_1, \dots, L_i, \dots, L_n) \succeq_i F(L_1, \dots, L_i^*, \dots, L_n)$

In the context of focal-set-oriented preferences, strategy-proofness means that if the outcome does not agree with an agent's labelling of its focal arguments, then the agent cannot alter this fact by mis-reporting its labelling. Formally, let  $L = F(L_1, \dots, L_i, \dots, L_n)$  be the aggregated labelling when  $i$  reports its own truthfully, and let  $L^* = F(L_1, \dots, L_i^*, \dots, L_n)$  be the aggregated result when  $i$  reports some arbitrary alternative  $L_i^*$ . Strategy-proofness means that  $\forall \alpha \in \bar{A}^i, L_i(\alpha) \neq L(\alpha)$  implies  $L_i(\alpha) \neq L^*(\alpha)$ .

THEOREM 1. Let  $I$  be a set of focal-set-oriented agents. The argument-wise plurality rule  $M(\cdot)$  is strategy-proof.

PROOF. Let  $i \in I$  be an arbitrary agent with labelling  $L_i$ . For convenience, let  $L = M(L_1, \dots, L_i, \dots, L_n)$  be the aggregated labelling when  $i$  reports its own truthfully (while others may or may not), and let  $L^* = M(L_1, \dots, L_i^*, \dots, L_n)$  be the aggregated result when  $i$  reports  $L_i^*$  (given the same  $L_{-i}$  reported by others). We need to prove that  $L_i(\alpha) \neq L(\alpha)$  implies  $L_i(\alpha) \neq L^*(\alpha)$  for all  $\alpha \in \bar{A}^i$ . We consider two cases:

(Case 1:  $L(\alpha) \in \{\text{in}, \text{out}, \text{undec}\}$ ): Suppose  $L_i(\alpha) \neq L(\alpha)$  for some  $\alpha \in \bar{A}^i$  (assumption). From the definition of the plurality rule, it follows that the plurality agreed on  $L(\alpha)$  as opposed to any other labelling of  $\alpha$ . Formally:

$$|\{j : L_j(\alpha) = L(\alpha)\}| > \max_{l \neq L(\alpha)} |\{j : L_j(\alpha) = l\}|$$

But since  $L_i(\alpha) \neq L(\alpha)$  it follows that, in particular, the plurality disagrees with  $i$ 's labelling of  $\alpha$ :

$$|\{j : L_j(\alpha) = L(\alpha)\}| > |\{j : L_j(\alpha) = L_i(\alpha)\}|$$

We now show that the above inequality does not change with any  $L_i^* \neq L_i$ . We have three possible cases:

1.  $L_i^*(\alpha) = L_i(\alpha)$  but  $L_i^*(\beta) \neq L_i(\beta)$  for one or more  $\beta \neq \alpha$  (be  $\beta \in \bar{A}^i$  or not). But this does not change the inequality, since  $M(L_1, \dots, L_i, \dots, L_n)[\alpha]$  is only dependent on votes on  $\alpha$ .
2.  $L_i^*(\alpha) = L(\alpha)$ , in which case the number of agents voting for  $L(\alpha)$  increases by 1, while the number of agents voting for  $L_i(\alpha)$  decreases by 1.
3.  $L_i^*(\alpha) = \{\text{in}, \text{out}, \text{undec}\} \setminus L(\alpha) \cup L_i(\alpha)$ , in which case the number of agents voting for  $L(\alpha)$  remains the same, while the number of agents voting for  $L_i(\alpha)$  decreases by 1.

<sup>2</sup>In the literature, strategy-proofness and incentive compatibility are sometimes used to mean the same thing, requiring us to state explicitly the type of equilibrium under which the mechanism is implemented (*e.g.* in dominant strategies).

Either way, the above inequality remains the same. Therefore, it will always be the case that  $L^*(\alpha) \neq L_i(\alpha)$  for  $\alpha \in \bar{A}^i$ .

(Case 2:  $L(\alpha) = \emptyset$ ): If for some  $\alpha \in \bar{A}^i$ ,  $L(\alpha) = \emptyset$ , then there exists three different labels,  $l_\alpha, l'_\alpha, l''_\alpha \in \{\text{in}, \text{out}, \text{undec}\}$  such that  $|\{j : L_j(\alpha) = l_\alpha\}| = |\{j : L_j(A) = l'_\alpha\}| \geq |\{j : L_j(A) = l''_\alpha\}|$ . Suppose that  $L_i(\alpha) = l_\alpha$  for  $\alpha \in \bar{A}^i$ . Then, if  $L_i^*(\alpha) = l'_\alpha$  or  $L_i^*(\alpha) = l''_\alpha$ , we still have  $M(L_1, \dots, L_i^*, \dots, L_n)[\alpha] \neq L_i(\alpha)$ .  $\square$

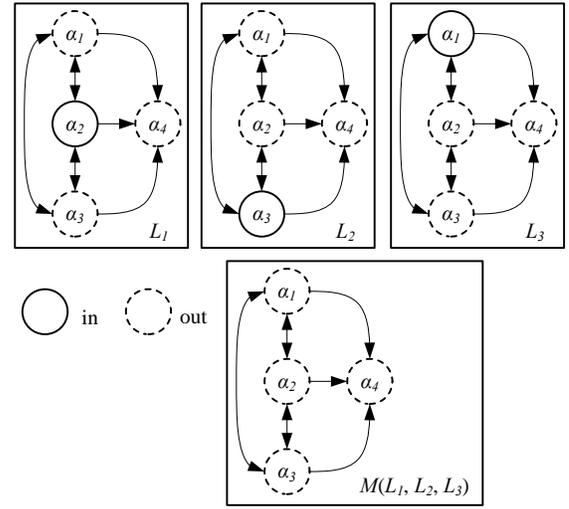
## 5.2 Other Social Choice Properties

Having analysed the strategic manipulability of argument-wise plurality voting, we now turn to analysing whether it satisfies the properties listed above.

**THEOREM 2.** *The argument-wise plurality voting operator  $M$  satisfies properties 1 to 7.*

**PROOF.** • **Universal Domain** follows immediately, since there are no restrictions on how each labelling  $L_i$  is defined.

- **Unanimity:** Consider an argument  $\alpha$  and a label  $l_\alpha$ . Since  $L_i(\alpha) = l_\alpha$  for each  $i$ , we have that  $|\{i : L_i(\alpha) = l_\alpha\}| = n$  and therefore  $M(L_1, \dots, L_n)[\alpha] = l_\alpha$ .
- **Anonymity:**  $M(L_1, \dots, L_n)[\alpha] = l_\alpha$  if and only if  $|\{i : L_i(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{i : L_i(\alpha) = l'_\alpha\}|$  if and only if  $|\{p(i) : L_{p(i)}(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{p(i) : L_{p(i)}(\alpha) = l'_\alpha\}|$ , which is equivalent to  $M(L_{p(1)}, \dots, L_{p(i)}, \dots, L_{p(n)})[\alpha] = l_\alpha$ .
- **Independence:** Suppose, towards a contradiction, that  $L_i(\alpha) = l_\alpha$  iff  $L'_i(\alpha) = l_\alpha$ , and  $M(L_1, \dots, L_n)[\alpha] = l_\alpha$  but  $M(L'_1, \dots, L'_n)[\alpha] \neq l_\alpha$ . Then  $|\{i : L_i(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{i : L_i(\alpha) = l'_\alpha\}|$ . But  $|\{i : L_i(\alpha) = l_\alpha\}| = |\{i : L'_i(\alpha) = l_\alpha\}|$  while for any  $l'_\alpha \neq l_\alpha$ ,  $|\{i : L_i(\alpha) = l'_\alpha\}| = |\{i : L'_i(\alpha) = l'_\alpha\}|$ . Then,  $|\{i : L'_i(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{i : L'_i(\alpha) = l'_\alpha\}|$ . Contradiction.
- **Systematicity:** Suppose (towards contradiction) that for some  $\alpha, \beta \in \mathcal{A}$ ,  $\forall i$ , we have  $L_i(\alpha) = L'_i(\beta)$ , and  $M(L_1, \dots, L_n)[\alpha] = l$ , but that  $M(L'_1, \dots, L'_n)[\beta] \neq l$ . Then  $|\{i : L_i(\alpha) = l\}| > \max_{l' \neq l} |\{i : L_i(\alpha) = l'\}|$ . But since  $|\{i : L_i(\alpha) = l\}| = |\{i : L'_i(\beta) = l\}|$ , we have  $|\{i : L'_i(\beta) = l\}| > \max_{l' \neq l} |\{i : L_i(\alpha) = l'\}|$ . And since  $|\{i : L_i(\alpha) = l'\}| = |\{i : L'_i(\beta) = l'\}|$  for all  $l' \neq l$ , we have  $|\{i : L'_i(\beta) = l\}| > \max_{l' \neq l} |\{i : L'_i(\beta) = l'\}|$ . This implies  $M(L'_1, \dots, L'_n)[\beta] = l$ . Contradiction.
- **Monotonicity:** Suppose that for  $\alpha \in \mathcal{A}$  and a label  $l_\alpha$  we have that  $L_i(\alpha) \notin l_\alpha$  while  $L'_i(\alpha) = l_\alpha$ , and we have that  $M(L_1, \dots, L_i, \dots, L_n)[\alpha] = l_\alpha$  while  $M(L_1, \dots, L'_i, \dots, L_n)[\alpha] \neq l_\alpha$ . But then,  $|\{j : L_j(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{j : L_j(\alpha) = l'_\alpha\}|$  in the profile  $(L_1, \dots, L_i, \dots, L_n)$  while in the profile  $(\hat{L}_1, \dots, \hat{L}_i, \dots, \hat{L}_n) \equiv (L_1, \dots, L'_i, \dots, L_n)$ , we have  $\{j : \hat{L}_j(\alpha) = l_\alpha\} = \{j : L_j(\alpha) = l_\alpha\} \cup \{i\}$ . Then  $|\{j : \hat{L}_j(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{j : \hat{L}_j(\alpha) = l'_\alpha\}|$ . Contradiction.
- **Non-dictatorship:** Suppose, towards a contradiction, that there exists an agent  $i$  such that for any profile  $(L_1, \dots, L_i, \dots, L_n)$ ,  $M(L_1, \dots, L_i, \dots, L_n) = L_i$ . Now consider  $\alpha \in \mathcal{A}$ , such that  $L_i(\alpha) = l_\alpha$  and  $\hat{L}$  such



**Figure 4: Counter example to collective rationality**

that  $\hat{L}(\alpha) \neq l_\alpha$ . Then,  $L_i(\alpha) \neq M(\hat{L}, \dots, L_i, \dots, \hat{L})$ . Contradiction.

$\square$

Despite these promising results, it turns out that plurality operator does not satisfy the collective rationality property.

**EXAMPLE 5.** Consider arguments  $\alpha_1, \alpha_2, \alpha_3$  and  $\alpha_4$ , with the attack relation depicted in Figure 4. Suppose we have three agents with the labellings  $L_1, L_2$  and  $L_3$ . We have:

- $M(L_1, L_2, L_3)[\alpha_1] = \text{out}$ ,
- $M(L_1, L_2, L_3)[\alpha_2] = \text{out}$ ,
- $M(L_1, L_2, L_3)[\alpha_3] = \text{out}$ ,
- $M(L_1, L_2, L_3)[\alpha_4] = \text{out}$ .

But then,  $M(L_1, L_2, L_3)$  is not a labelling (see 4).

The above counter-example is a variant of the discursive dilemma [10] in the context of argument evaluation, which itself is a variant of the well-known Condorcet paradox.

It is worth noting that, when the preferences are focal-set oriented, labellings are partitioned in two classes: top labellings, which satisfy the focal-set assignment of labels and bottom labellings, which do not. These kinds of preferences are called *dichotomous*. Brams and Fishburn [2] showed that *approval voting*, a method according to which each voter can vote for as many candidates as she likes, is strategy-proof on dichotomous preferences. So why not apply *labelling-wise approval voting* instead of argument-wise plurality?

As it turns out, approval voting on labellings also fails to satisfy collective rationality. Just consider a system with only two arguments,  $\alpha$  and  $\beta$  in a cycle of mutual defeat. Three rational labellings are possible for  $(\alpha, \beta)$ : **(in, out)**, **(out, in)** and **(undec, undec)**. Suppose there are only two agents, 1 and 2, with focal sets,  $\{\alpha\}$  and  $\{\beta\}$ , respectively. Then each one will have a top preferred labelling, **(in, out)** for 1 and **(out, in)** for 2. Each one will vote only for her top labelling. Then, instead of having a single labelling as an outcome, a tie obtains, i.e. a set of two labellings, which certainly is not a rational labelling.

## 6. IS ‘GOOD’ AGGREGATION POSSIBLE?

In the previous section, we analysed a particular judgement aggregation operator (namely, argument-wise plurality voting). We showed that while it satisfies most key properties, it fails to always generate collectively rational judgements. This is a significant limitation, and gives rise to a more important question of whether any such operator exists. We now give a negative answer to this question, then show how this impossibility result can be avoided by restricting the domain of the voting rule.

### 6.1 An Impossibility Result

Social Choice Theory has been built around an impossibility result on the aggregation of preferences (Arrow’s Theorem). A similar result has been found on the aggregation of judgements in propositional settings [10] and extended to more general logics [7]. The theorem below provides a counter-part for abstract argumentation framework.

We now show that there exists no possible aggregation operator  $F$  that satisfies collective rationality along with only four other minimal conditions, namely: universal domain, anonymity, systematicity, and unanimity.

**THEOREM 3.** *There exists no  $F$  satisfying Universal Domain, Anonymity, Systematicity, Unanimity, and Collective Rationality.*

**PROOF.** *By Universal Domain we can consider any profile of labellings. Let us just focus on two profiles  $(L_1, \dots, L_n)$  and  $(L'_1, \dots, L'_n)$  such that there exists a permutation  $p : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$  such that  $L'_i = L_{p(i)}$  for each  $i$ . By Anonymity  $F$  must yield the same outcomes over both profiles over any particular argument.*

*By systematicity, we have the following implication:*

$$\forall \alpha, \beta \in \mathcal{A} \text{ if } \forall i, L_i(\alpha) = L'_i(\beta) \\ \text{then } F(L_1, \dots, L_n)[\alpha] = F(L'_1, \dots, L'_n)[\beta] \quad (1)$$

*which can be rewritten as (with  $LAB = \{\mathbf{in}, \mathbf{out}, \mathbf{undec}\}$ ):*

$$\forall \alpha, \beta \in \mathcal{A} \text{ if } \forall l \in LAB, \{i : L_i(\alpha) = l\} = \{i : L'_i(\beta) = l\} \\ \text{then } F(L_1, \dots, L_n)[\alpha] = F(L'_1, \dots, L'_n)[\beta] \quad (2)$$

*Moreover, by Anonymity we have:*

$$F(L'_1, \dots, L'_n)[\beta] = F(L_1, \dots, L_n)[\beta]$$

*So, the above implication can be rewritten again as:*

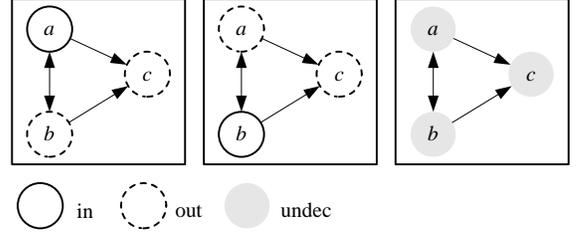
$$\forall \alpha, \beta \in \mathcal{A} \text{ if } \forall l \in LAB, \{i : L_i(\alpha) = l\} = \{i : L'_i(\beta) = l\} \\ \text{then } F(L_1, \dots, L_n)[\alpha] = F(L_1, \dots, L_n)[\beta] \quad (3)$$

*Now, by Universal Domain consider without loss of generality an argumentation framework  $AF = \langle \{a, b, c\}, \rightarrow \rangle$  with  $\rightarrow = \{(a, b), (b, a), (a, c), (b, c)\}$ . This graph is shown in the figure below, with all its three possible labellings. By Universal Domain, assume there is an even number  $n$  of agents. Now, define profile  $(L_1, \dots, L_n)$  such that:*

- $L_i(a) = \mathbf{out}$ ,  $L_i(b) = \mathbf{in}$ , and  $L_i(c) = \mathbf{out}$  for  $i = 1, \dots, \frac{n}{2}$ ;
- $L_i(a) = \mathbf{in}$ ,  $L_i(b) = \mathbf{out}$ , and  $L_i(c) = \mathbf{out}$  for  $i = \frac{n}{2} + 1, \dots, n$ .

*Both these cases correspond to legitimate labellings.*

*Now, define profile  $(L'_1, \dots, L'_n)$  such that*



- $L'_i(a) = \mathbf{in}$ ,  $L'_i(b) = \mathbf{out}$ , and  $L'_i(c) = \mathbf{out}$  for  $i = 1, \dots, \frac{n}{2}$ ;
- $L'_i(a) = \mathbf{out}$ ,  $L'_i(b) = \mathbf{in}$ , and  $L'_i(c) = \mathbf{out}$  for  $i = \frac{n}{2} + 1, \dots, n$ .

*Clearly,  $(L'_1, \dots, L'_n)$  is a permutation of  $(L_1, \dots, L_n)$ . And we have:*

$$\{i : L_i(a) = \mathbf{out}\} = \{i : L'_i(b) = \mathbf{out}\} = \{1, \dots, \frac{n}{2}\} \quad (4)$$

*Likewise, we have:*

$$\{i : L_i(a) = \mathbf{in}\} = \{i : L'_i(b) = \mathbf{in}\} = \{\frac{n}{2} + 1, \dots, n\} \quad (5)$$

*From formulas 5 and 4, we have:*

$$\forall l \in LAB, \{i : L_i(a) = l\} = \{i : L'_i(b) = l\} \quad (6)$$

*By formulas 3 and 6, it follows that:*

$$F(L_1, \dots, L_n)[a] = F(L_1, \dots, L_n)[b] \quad (7)$$

*Now consider  $F(L_1, \dots, L_n)$ . Note that  $a$  and  $b$  cannot be both  $\mathbf{in}$  or both  $\mathbf{out}$ , since this would violate collective rationality. The only outcome consistent with formula 7 is  $F(L_1, \dots, L_n)[a] = \mathbf{undec}$  and  $F(L_1, \dots, L_n)[b] = \mathbf{undec}$ . However, as a consequence of collective rationality, this would yield  $F(L_1, \dots, L_n)[c] = \mathbf{undec}$ , which violates unanimity (because  $\forall i, L_i(c) = \mathbf{out}$  and therefore unanimity requires that  $F(L_1, \dots, L_n)[c] = \mathbf{out}$ ). Therefore, one cannot simultaneously satisfy universal domain, anonymity, systematicity, and unanimity.  $\square$*

The proof is, in words, as follows. Given any profile and its permutation, any supposed aggregation operator yields the same result. By systematicity the result should be obtained on the same profile (recall that the other just obtains by permutating the names of the agents) in the same way for any pair of arguments. But this means that the number of agents that vote for the ‘winning’ labellings on both arguments must be the same. Then a profile is constructed, for which the aggregation operator is unable to yield an outcome without violating this last requirement.

The above impossibility result highlights a major barrier to reaching good collective argument evaluation in general. This is similar in flavour to List and Pettit’s impossibility theorem on judgement aggregation in propositional logic [10] (with the addition of unanimity). In our context, the result means that rational aggregation on argument evaluation can only be achieved at a cost to universal domain, unanimity, anonymity or systematicity.

### 6.2 Circumventing the Impossibility

Faced with the impossibility result, how can agents guarantee, in some way, reaching collective argument evaluation

that is collectively rational? Following the tradition of social choice theory, we explore what restrictions on the domain of the argument-wise plurality voting rule guarantee collective rationality. In particular, we provide a full characterisation of the space of labelling profiles that guarantees collective rationality using the argument-wise plurality voting rule.

We first need a few conditions. The first is the *no-tie* condition which, as the name suggests, means that agents can always make a collective decision on each argument.

**DEFINITION 9 (NO-TIE).** *Labelling profile*  $(L_1, \dots, L_n)$  *satisfies the no-tie condition if for any*  $\alpha \in \mathcal{A}$ , *there exists a label*  $l$  *such that*  $|\{i : L_i(\alpha) = l\}| > \max_{l' \neq l} |\{i : L_i(\alpha) = l'\}|$ .

Next, we present the notion of Condorcet winner, which captures the plurality winner on an individual argument.

**DEFINITION 10 (CONDORCET WINNER).** *We say that a label*  $l_\alpha \in \{\mathbf{in}, \mathbf{out}, \mathbf{undec}\}$  *of an argument*  $\alpha \in \mathcal{A}$  *is a Condorcet Winner with respect to a labelling profile*  $(L_i)_{i=1}^n$ , *denoted*  $CW(\alpha, l_\alpha, (L_i)_{i=1}^n)$  *iff*  $|\{i : L_i(\alpha) = l_\alpha\}| > |\{i : L_i = l'_\alpha\}|$  *for every label*  $l'_\alpha \neq l_\alpha$ .

Next, we list the *Condorcet defeat* condition. Intuitively, this condition means that if an argument  $\alpha$  is collectively rejected by the agents, then the agents must also collectively agree (via plurality) on accepting at least one of the counter-arguments against  $\alpha$ . In other words, the agents' individual attacks on  $\alpha$  are not arbitrary, but must exhibit some minimal degree of consensus.

**DEFINITION 11 (CONDORCET DEFEAT).** *A labelling profile*  $(L_1, \dots, L_n)$  *satisfies Condorcet defeat if and only if the following holds:*

$CW(\alpha, \mathbf{out}, (L_i)_{i=1}^n)$  *if and only if*  $\exists \beta \in \mathcal{A}$ , *such that*  $\beta \rightarrow \alpha$  *and*  $CW(\beta, \mathbf{in}, (L_i)_{i=1}^n)$ .

Finally, we need the following condition, which we call *non-Condorcet indecision*. Intuitively, it requires that if an argument  $\alpha$  is collectively accepted by the agents, then the agents must never collectively be undecided on or accept any of the counter-arguments against  $\alpha$ . Notice that, unlike the existence condition in Condorcet defeat, here the lack of indecision must hold for all defeaters.

**DEFINITION 12 (NON-CONDORCET INDECISION).** *A labelling profile satisfies non-Condorcet indecision if and only if the following holds:*

$CW(\alpha, \mathbf{in}, (L_i)_{i=1}^n)$  *if and only if*  $\nexists \beta \in \mathcal{A}$ , *such that*  $\beta \rightarrow \alpha$  *satisfies either*  $CW(\beta, \mathbf{undec}, (L_i)_{i=1}^n)$  *or*  $CW(\beta, \mathbf{in}, (L_i)_{i=1}^n)$ .

We are now ready to define the necessary and sufficient restrictions on labelling profiles that guarantee collective rationality under argument-wise plurality voting.

**THEOREM 4.** *The argument-wise plurality voting rule*  $M$  *satisfies collective rationality if and only if each labelling profile*  $(L_1, \dots, L_n)$  *in its domain satisfies Condorcet Defeat, Non-Condorcet Indecision and the No-Tie condition.*

**PROOF.**  $\Leftarrow$ : *Let*  $L = (L_1, \dots, L_n)$  *be an arbitrary profile in the domain of*  $F$  *satisfying coordinated collective defeat. By the No-Tie condition, for any profile*  $M(L_1, \dots, L_n)[\alpha] \in \{\mathbf{in}, \mathbf{out}, \mathbf{undec}\}$ . *We now need to show that*  $F$  *satisfies the conditions of legal labelling [3].*

- *Suppose*  $M(L_1, \dots, L_n)[\alpha] = \mathbf{in}$ . *We want to prove that*  $\forall \beta$  *such that*  $\beta \rightarrow \alpha$ , *we have*  $M(L_1, \dots, L_n)[\beta] = \mathbf{out}$ . *By hypothesis,*  $|\{i : L_i(\alpha) = \mathbf{in}\}| > \max_{l'_\alpha \neq \mathbf{in}} |\{i : L_i(\alpha) = l'_\alpha\}|$ . *In particular:*

$$|\{i : L_i(\alpha) = \mathbf{in}\}| > |\{i : L_i(\alpha) = \mathbf{out}\}| \quad (8)$$

*But since each*  $L_i$  *is a labelling, we have that*  $\forall i$ , *if*  $L_i(\alpha) = \mathbf{in}$  *then for every*  $\beta \rightarrow \alpha$ ,  $L_i(\beta) = \mathbf{out}$ . *That is:*

$$|\{i : L_i(\alpha) = \mathbf{in}\}| \leq |\{i : L_i(\beta) = \mathbf{out}\}| \quad (9)$$

*Now let us suppose (by contradiction) that for some*  $\beta \rightarrow \alpha$ ,  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ . *Then we have*  $|\{i : L_i(\beta) = \mathbf{in}\}| > \max_{l'_\beta \neq \mathbf{in}} |\{i : L_i(\beta) = l'_\beta\}|$ . *Hence, in particular:*

$$|\{i : L_i(\beta) = \mathbf{in}\}| > |\{i : L_i(\beta) = \mathbf{out}\}| \quad (10)$$

*By transitivity (from Equations 9 and 10) we have:*

$$|\{i : L_i(\beta) = \mathbf{in}\}| > |\{i : L_i(\alpha) = \mathbf{in}\}| \quad (11)$$

*But on the other hand, by definition of labelling, for each*  $i$  *such that*  $L_i(\beta) = \mathbf{in}$  *we have that*  $L_i(\alpha) = \mathbf{out}$ . *This means that*  $\{i : L_i(\beta) = \mathbf{in}\} \subseteq \{i : L_i(\alpha) = \mathbf{out}\}$ , *which in turn implies that*

$$|\{i : L_i(\beta) = \mathbf{in}\}| \leq |\{i : L_i(\alpha) = \mathbf{out}\}| \quad (12)$$

*But then (from Equations 11 and 12) we get:*

$$\begin{aligned} |\{i : L_i(\alpha) = \mathbf{out}\}| &\geq |\{i : L_i(\beta) = \mathbf{in}\}| \\ &> |\{i : L_i(\alpha) = \mathbf{in}\}| \end{aligned} \quad (13)$$

*Contradiction (between Equations 8 and 13).*

*Suppose, on the other hand, that for some*  $\beta \rightarrow \alpha$ ,  $M(L_1, \dots, L_n)[\beta] = \mathbf{undec}$ . *This cannot hold since it contradicts the non-Condorcet indecision assumption.*

*Thus,  $\forall \beta \in \mathcal{A}$ , if*  $\beta \rightarrow \alpha$ , *we have*  $M(L_1, \dots, L_n)[\beta] = \mathbf{out}$ , *and the labelling condition is satisfied.*

- *Suppose, on the other hand, that*  $M(L_1, \dots, L_n)[\alpha] = \mathbf{out}$ . *We want to prove that*  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ , *for some*  $\beta \rightarrow \alpha$ . *By Condorcet Defeat,  $\exists \beta \in \mathcal{A}$ ,  $\beta \rightarrow \alpha$ , such that*  $|\{i : L_i(\beta) = \mathbf{in}\}| > \max_{l'_\beta \neq \mathbf{in}} |\{i : L_i(\beta) = l'_\beta\}|$ . *That is,*  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ .
- *Finally, let us suppose that*  $M(L_1, \dots, L_n)[\alpha] = \mathbf{undec}$ . *We want to prove that*  $\exists \beta \in \mathcal{A}$  *such that*  $\beta \rightarrow \alpha$  *and*  $M(L_1, \dots, L_n)[\beta] = \mathbf{undec}$ . *Since*  $M(L_1, \dots, L_n)[\alpha] = \mathbf{undec}$ ,  $M(L_1, \dots, L_n)[\alpha] \neq \mathbf{in}$ . *By non-Condorcet Indecision we have that there exists*  $\beta \rightarrow \alpha$  *such that either*  $M(L_1, \dots, L_n)[\beta] = \mathbf{undec}$  *or*  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ . *But since*  $M(L_1, \dots, L_n)[\alpha] \neq \mathbf{out}$ , *by Condorcet defeat we have*  $\forall \beta$  *such that*  $\beta \rightarrow \alpha$ ,  $M(L_1, \dots, L_n)[\beta] \neq \mathbf{in}$ . *Then,*  $M(L_1, \dots, L_n)[\beta] = \mathbf{undec}$ .

$\Rightarrow$ : *Suppose that*  $M$  *satisfies the Collective Rationality condition. By definition each profile in its domain satisfies the No-tie condition. To see that it satisfies also the Condorcet defeat and non-Condorcet indecision condition, just consider for every*  $\alpha \in \mathcal{A}$  *two possible cases:*

- *if*  $M(L_1, \dots, L_n)[\alpha] = \mathbf{out}$ , *then by definition of Collective Rationality,  $\exists \beta \in \mathcal{A}$ , with*  $\beta \rightarrow \alpha$  *such that*  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ . *But this means that*  $|\{i :$

$L_i(\beta) = \mathbf{in}\} > |\{i : L_i(\beta) = l'_\beta\}|$  for every  $l'_\beta \neq \mathbf{in}$ . But then,  $CW(\beta, \mathbf{in}, (L_i)_{i=1}^n)$ . Conversely, if there exists  $\beta \rightarrow \alpha$  such that  $|\{i : L_i(\beta) = \mathbf{in}\}| > |\{i : L_i(\beta) = l'_\beta\}|$  for every  $l'_\beta \neq \mathbf{in}$ , we have that  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ . Then, by Collective Rationality,  $M(L_1, \dots, L_n)[\alpha] = \mathbf{out}$ . Thus, Condorcet defeat is satisfied.

- If  $M(L_1, \dots, L_n)[\alpha] = \mathbf{in}$  for some  $\alpha \in \mathcal{A}$ , then by definition of Collective Rationality  $\forall \beta \in \mathcal{A}$  with  $\beta \rightarrow \alpha$ ,  $M(L_1, \dots, L_n)[\beta] = \mathbf{out}$ . But this trivially means that  $\nexists \beta' \in \mathcal{A}$ ,  $\beta' \rightarrow \alpha$  such that either  $M(L_1, \dots, L_n)[\beta'] = \mathbf{undec}$  or  $M(L_1, \dots, L_n)[\beta'] = \mathbf{in}$ . Conversely, suppose  $\nexists \beta \in \mathcal{A}$  with  $\beta \rightarrow \alpha$  such that  $M(L_1, \dots, L_n)[\beta] = \mathbf{undec}$  or  $M(L_1, \dots, L_n)[\beta] = \mathbf{in}$ . Then, for every  $\beta \rightarrow \alpha$ , we have  $M(L_1, \dots, L_n)[\beta] = \mathbf{out}$ . But then, by Collective Rationality,  $M(L_1, \dots, L_n)[\alpha] = \mathbf{in}$ . Thus, non-Condorcet indecision is always satisfied.

□

The careful reader will notice that the conditions of Condorcet Defeat and non-Condorcet Indecision, required for the result, actually correspond to the requirements of well-defined labellings (recall Definition 5). Indeed, this shows that collective rationality requires strong conditions on the collective structure of agents' labellings. These conditions are quite strong, in the sense that they cannot be reduced to properties of the individual labellings.

The full characterisation provided above has another consequence. In order to achieve collective rationality while only appealing to restrictions on *individual* labellings, we would need to make even stronger assumptions to those in Theorem 4. For example, we could require that whenever an agent labels an argument as **out**, then it must label each of its defeaters as **in**, and so on. While these kinds of restrictions guarantee the necessary partial consensus among agents, they are extremely unrealistic (even less realistic than the ones shown in the theorem). This reveals that satisfying collective rationality is not easily achievable in practice with a kind of argument-wise plurality vote.

## 7. CONCLUSION & RELATED WORK

In this paper, we explored the following question: *Given an argument structure and a set of agents, each with a legitimate subjective evaluation of the given arguments, how can the agents reach a collective compromise on the evaluation of those arguments?* We presented three important results: (1) proved that argument-wise plurality voting satisfies many well-known social choice properties, albeit not collective rationality; (2) proved the impossibility of any aggregation operator that simultaneously satisfies collective rationality together with universal domain, unanimity, anonymity and systematicity; and (3) fully characterised the space of individual judgements that guarantees collective rationality using argument-wise plurality voting.

Recently, Caminada and Pigozzi [4] presented some operators for aggregating multiple argument labellings into a single labelling. They focused on a 'compatibility' property: *that the social outcome must not go against any individual judgement*, and showed that this can be achieved together with collective rationality. However, they did not explore whether these operators could satisfy other classical social-choice properties. Our results provide an important comple-

ment to their work, by identifying bounds on what can be achieved simultaneously by *any* aggregation operator.

Coste-Marquis et al explored the problem of aggregating multiple argumentation frameworks [6]. However, each agent contributes a different argument graph altogether, rather than a judgement of how a given graph must be evaluated.

We have only scratched the surface of the intersection between 'judgement aggregation' and collective argument evaluation. A key research agenda is studying which 'judgement aggregation' results carry over, and how this is affected by the structure of the collective argument evaluation problem.

## Acknowledgments

We are grateful to Richard Booth and the anonymous reviewers for their insightful comments. We are particularly grateful to Martin Caminada for helping us strengthen the technical results.

## 8. REFERENCES

- [1] K. J. Arrow. *Social choice and individual values*. Wiley, New York NY, USA, 1951.
- [2] S. Brams and P. Fishburn. Approval voting. *American Political Science Review*, 72(3):831–847, 1978.
- [3] M. Caminada and D. M. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2-3):109–145, 2009.
- [4] M. Caminada and G. Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, (to appear).
- [5] M. W. A. Caminada. On the issue of reinstatement in argumentation. In *Proc. European Conference on Logics in Artificial Intelligence (JELIA)*, volume 4160 of *LNCS*, pages 111–123. Springer, 2006.
- [6] S. Coste-Marquis, C. Devred, S. Konieczny, M.-C. Lagasque-Schiex, and P. Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171(10–15):730–753, 2007.
- [7] F. Dietrich. A generalised model of judgment aggregation. *Social Choice and Welfare*, 28:529–565, 2007.
- [8] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.
- [9] W. Gärtner. *A Primer on Social Choice Theory*. Oxford University Press, 2006.
- [10] C. List and P. Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18:89–110, 2002.
- [11] C. List and C. Puppe. Judgment aggregation: a survey. In *The Oxford handbook of rational and social choice*. Oxford University Press, Oxford, UK, 2009.
- [12] A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, New York NY, USA, 1995.