Experimental Assessment of Aggregation Principles in Argumentation-enabled Collective Intelligence

EDMOND AWAD, The Media Lab, Massachusetts Institute of Technology and Masdar Institute
JEAN-FRANÇOIS BONNEFON, Toulouse School of Economics, Center for Research in Management, Institute for Advanced Study in Toulouse, University of Toulouse Capitole
MARTIN CAMINADA, School of Computer Science & Informatics, Cardiff University
THOMAS MALONE, Sloan School of Management, Massachusetts Institute of Technology
IYAD RAHWAN*, The Media Lab, Massachusetts Institute of Technology and Masdar Institute

On the Web, there is always a need to aggregate opinions from the crowd (as in posts, social networks, forums, etc.). Different mechanisms have been implemented to capture these opinions such as fiLikefi in Facebook, fiFavoritefi in Twitter, thumbs-up/down, flagging, and so on. However, in more contested domains (e.g. Wikipedia, political discussion, and climate change discussion) these mechanisms are not sufficient since they only deal with each issue independently without considering the relationships between different claims. We can view a set of conflicting arguments as a graph in which the nodes represent arguments and the arcs between these nodes represent the defeat relation. A group of people can then collectively evaluate such graphs. To do this, the group must use a rule to aggregate their individual opinions about the entire argument graph. Here, we present the first experimental evaluation of different principles commonly employed by aggregation rules presented in the literature. We use randomized controlled experiments to investigate which principles people consider better at aggregating opinions under different conditions. Our analysis reveals a number of factors, not captured by traditional formal models, that play an important role in determining the efficacy of aggregation. These results help bring formal models of argumentation closer to real-world application.

CCS Concepts: **•Computing methodologies →Nonmonotonic, default reasoning and belief revision; Multi-agent systems; •Applied computing →Law, social and behavioral science;**

Additional Key Words and Phrases: Argumentation, Voting, Experiment

ACM Reference format:
Edmond Awad, Jean-François Bonnefon, Martin Caminada, Thomas Malone, and Iyad Rahwan. 2017. Experimental Assessment of Aggregation Principles in Argumentation-enabled Collective Intelligence. *ACM Trans. Internet Technol.* 0, 0, Article 0 (February 2017), 20 pages.
DOI: 0000001.0000001

*The corresponding author

Author's addresses: E. Awad, The Media Lab, Massachusetts Institute of Technology, USA; Masdar Institute, UAE, Email: awad@mit.edu; Jean-François Bonnefon, Toulouse School of Economics, Center for Research in Management, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, France, Email: jean-francois.bonnefon@tse-fr.eu; Martin Caminada, School of Computer Science & Informatics, Cardiff University, UK, Email: CaminadaM@cardiff.ac.uk; Thomas Malone, Sloan School of Management, Massachusetts Institute of Technology, USA, Email: malone@mit.edu; Iyad Rahwan, The Media Lab, Massachusetts Institute of Technology, USA; Masdar Institute, UAE, Email: irahwan@mit.edu.

1 INTRODUCTION

In many online systems that support *Collective Intelligence* (e.g. Wikipedia, question answering systems, and discussion forums) different conflicting points of view arise, even based on the same information (Apic et al. 2011; Introne et al. 2011; Marvel et al. 2011). This raises the challenge of supporting consensus-making among community members. However, many systems on the Web employ arbitrary aggregation rules to handle such tasks, due to the lack of clear appropriate rule given the settings at hand, and the difficulty of evaluating the potential alternatives. Further, the scalability, and the complexity of consensus-making in online systems makes the use of traditional voting rules inadequate.

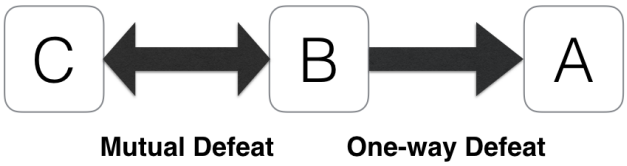
A crucial step towards applying online concensus-making is the representation of information. Argumentation has been shown to provide a realistic environment to represent the conflict on the Web (Buckingham Shum 2008; Klein and Iandoli 2008; Rahwan et al. 2007). One of the most influential frameworks is probably the abstract argumentation framework, which was proposed by Dung (Dung 1995). In this framework, arguments are repesented by nodes, and the defeat relations between these arguments are represented by arcs, which form an argumentation graph. This argumentation graph can be evaluated (that is, some arguments are accepted, and others are rejected) in multiple consistent ways. In multi-agent settings, where different agents can subscribe to different evaluations, an aggregation rule is needed to produce a collective consistent evaluation. Unlike the case with simple aggregation of opinions for isolated propositions, aggregation in the context of argumentation can pose further restrictions to ensure consistency. Consider the following example, which considers making judgments when conflicting information is provided.

Example 1.1 (Suspect Stephen). A committee of 10 jury members was formed to make a collective decision about whether there is evidence against Stephen, a potential suspect. The committee is provided with the arguments that were laid down by the opposing sides, to help them make an informed decision. Understandably, arguments of one side are in conflict with the other side's arguments (refer to Figure 1 (a)). The relations among arguments, are represented by an argumentation graph, in which nodes are arguments, and arcs are the defeat relations between arguments (as in Figure 1 (a))).

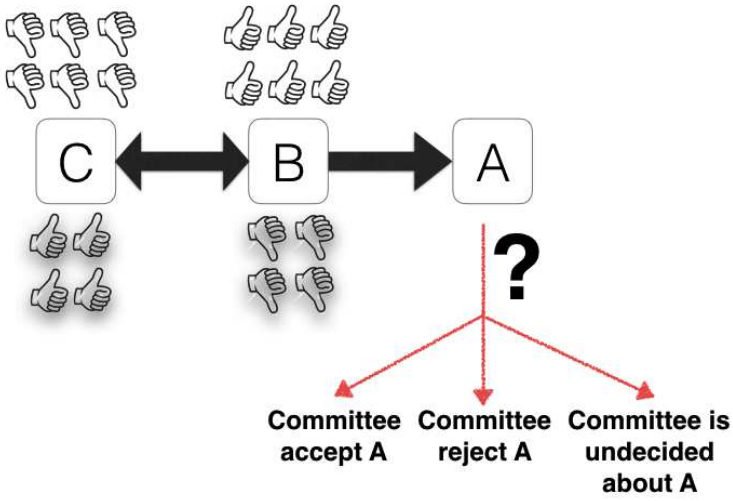
Each of the ten members is expected to have a reasonable judgement about which arguments should be accepted. Suppose that six out of the ten members believe that argument B should be accepted, and argument C should be rejected (thumbs-up/down above graph in Figure 1 (b)), while the other four think that argument C should be accepted and argument B should be rejected (thumbs-up/down below graph). The question is the following: should the committee collectively accept argument A, thus accepting there is an evidence against Stephen, or not?

The problem highlighted in the previous example encompasses two issues: First, the provided information is inconsistent. For example, one can see that arguments B and C cannot be accepted together. Likewise, the evaluation of argument B influences that of argument A i.e. if B is accepted, then A should be rejected, and vice versa. As such, the evaluation of argument A is influenced by the evaluation of B and C. Aggregating judgments over such a set of inconsistent information is not as simple as it seems. Since in this example argument A contains the conclusion of interest, one can suggest two possible methods as to how collectively evaluate A given indviduals' evaluation: 1) ignore the individuals' evaluation of arguments B and C and only aggregate evaluations of argument A, and 2) ignore the individuals' evaluation of argument A, and aggregate evaluation of arguments B and C, and then use the outcome to determine the collective evaluation of argument A. Applying these two methods on this example provides a similar

A: The witness saw Stephen in the parking area next to the crime scene. Therefore, there is evidence against Stephen.
B: It was dark then. Therefore, the witness probably mistook someone else for Stephen.
C: The parking area is well lit. Therefore, the witness could clearly identify Stephen.



(a)



(b)

Fig. 1. **An example of aggregation in contested domains.**

outcome. Unfortunately, these two methods do not always yield the same outcome, as was shown in (Awad et al. 2015).[1] Further, choosing between these two methods is not a simple task. There have been many studies discussing the choice

[1]In fact, this problem mirrors the *discursive dilemma* (Pettit 2001) in judgement aggregation (JA). The *discursive dilemma* refers to the paradox that the use of the "simple" majority rule to aggregate "consistent" individuals' judgments of logically related issues can result into "inconsistent" collective judgment of these issues.

between conclusion-based and premise-based methods, the counterparts of the above suggested methods (1 and 2, respectively) in other fields of aggregation like judgement aggregation (JA) (Bonnefon 2010; Bovens and Rabinowicz 2006; Pettit 2001; Pigozzi 2006). This suggests the inadequacy of classical voting in handling such problems.

The second issue is concerned with the amount of support needed to collectively make a judgment, and it mirrors the issue of choosing among the supermajority rules with the *majority* and the *unanimity* rules being the two extremes. A *supermajority* rule requires that for an alternative to be chosen collectively it has to receive at least k votes, where $k > 0.5 \times N$ and N is the number of voters ($k = N$ for *unanimity* and $k = \lfloor 1 + 0.5 \times N \rfloor$ for *majority*). In the above example, using the majority rule would result in the group concluding that *there is no evidence against Stephen*, while using unanimity would result in the group being undecided about whether *there is evidence against Stephen or not*. Although the comparison between these two methods has been studied extensively, both formally and experimentally (Guttman 1998; Miller and Vanberg 2015; Quesada 2011), it has never been studied experimentally in contested domains where conflicting information are considered in making a collective decision.

Recently, various rules were proposed to appropriately aggregate evaluations of argumentation graphs. The argument-wise plurality rule (AWPR), which chooses the collective evaluation of each argument by plurality was defined and analyzed in (Awad et al. 2015; Rahwan and Tohmé 2010). On the other hand, Caminada and Pigozzi (Caminada and Pigozzi 2011) showed how judgment aggregation concepts can be applied to formal argumentation in a different way. They proposed three possible operators for aggregating labellings, namely the sceptical operator, the credulous operator, and the super credulous operator (collectively shortened here as SSCOs). These operators guarantee not only a well-formed outcome but also a compatible one, that is, it does not go against the judgment of any individual. The analysis of the above methods so far has been restricted to a principle-based approach such as the one devised by Arrow (Arrow 1951; Arrow et al. 2002), by evaluating aggregation rules on the base of satisfying seemingly plausible, "fairness" postulates.

In this study, we offer the first experimental-based study of aggregation rules in contested domains. We experimentally compare the desirability of the principles employed by AWPR and SSCOs. We find that principles employed by AWPR are more favorable. However, there can be some conditions in which this is not the case. Our results provide clear suggestions in regard of what aggregation rules to use given some assumed contexts, and offer a first step towards the evaluation of aggregation rules in contested domains.

2 THEORETICAL BACKGROUND: ABSTRACT ARGUMENTATION AND LABELING AGGREGATION RULES

In order to form reasonable judgments on arguments, a formal representation of the relations is needed. Arguments and the relationship between them can be represented as a directed graph in which nodes represent arguments, and arcs between nodes represent binary *defeat* relations over them. This framework is known as abstract argumentation framework, proposed by Dung (Dung 1995). In the previous example, one can see that arguments B and C defeat each other, and argument B defeats argument A.

Definition 2.1 (Argumentation framework). An *argumentation framework* is a pair $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ where $\mathcal{A}$ is a finite set of arguments and $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ is a defeat relation. An argument a *defeats* an argument b if $(a, b) \in \rightharpoonup$ (sometimes written $a \rightharpoonup b$).

For example, in Figure 2, argument a_1 is defeated by arguments a_2 and a_4 which are, in turn, defeated by arguments a_3 and a_5.

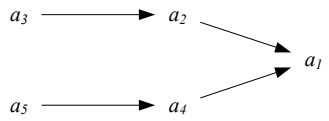


Fig. 2. A simple argument graph

One then can evaluate each argument (i.e. label (Caminada 2006; Caminada and Gabbay 2009) each argument) by accepting it (i.e. labeling it as in), rejecting it (i.e. labeling it as out), or being undecided about it (i.e. labeling it as undec). Formally, a labeling is a total function:

Definition 2.2 (Argument labeling). Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. An *argument labeling* is a total function $L : \mathcal{A} \to \{\texttt{in}, \texttt{out}, \texttt{undec}\}$.

This evaluation (or labeling), however, should follow some rules. One of the essential properties, that is common, is the condition of *completeness*, and is captured, in terms of labelings, in the following two conditions:

(1) An argument is labeled *accepted* (or *in*) if and only if all its defeaters are rejected (or *out*).
(2) An argument is labeled *rejected* (or *out*) if and only if at least one of its defeaters is accepted (or *in*).

In all other cases, an argument should be labeled *undecided* (or *undec*). Thus, evaluating a set of arguments amounts to labeling each argument using a labeling function $\mathcal{L} : \mathcal{A} \to \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ to capture these three possible labels. Any labeling that satisfies the above conditions is a *legal labeling*, or a *complete labeling*. These conditions can equivalently be formulated as follows.

Definition 2.3 (Complete labeling). Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. A *complete labeling* is a total function $L : \mathcal{A} \to \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ such that:

- $\forall a \in \mathcal{A}$: if $L(a) = \texttt{in}$ then $\forall b \in \mathcal{A}$: $(b \rightharpoonup a \implies L(b) = \texttt{out})$;
- $\forall a \in \mathcal{A}$: if $L(a) = \texttt{out}$ then $\exists b \in \mathcal{A}$ s.t. $(b \rightharpoonup a \wedge L(b) = \texttt{in})$; and
- $\forall a \in \mathcal{A}$: if $L(a) = \texttt{undec}$ then
  - $\exists b \in \mathcal{A}$: $(b \rightharpoonup a \wedge L(b) = \texttt{undec})$; and
  - $\nexists b \in \mathcal{A}$: $(b \rightharpoonup a \wedge L(b) = \texttt{in})$

From Example 1.1, one can see that there can be different reasonable positions regarding the evaluation of an argumentation graph (following the previous conditions). Thus, choosing a legal labeling above another becomes a matter of preference. Therefore, in a multi-agent setting, there has been some aggregation rules that were proposed to aggregate preferences over labelings. The work (Awad et al. 2015; Rahwan and Tohmé 2010) defined and analyzed the argument-wise plurality rule (AWPR) which chooses the collective label of each argument by plurality, independently from other arguments (i.e. for each argument, among in, out, and undec, the one with the most votes is chosen as the collective label for this argument). On the other hand, Caminada and Pigozzi (Caminada and Pigozzi 2011) proposed three operators for aggregating labelings, namely the sceptical operator, the credulous operator, and the super credulous operator (collectively shortened as SSCOs). At the crux of these operators is the notion that an argument is not collectively accepted or collectively rejected unless this decision receives a unanimous support. The level of this unanimously required "support" though varies across the three rules. While one of the rules defines "support" as a strict agreement by everyone to accept (or reject) an argument in order to collectively accept (or reject) this argument, the

other two rules are more lenient regarding the support requirement, as they require only some voters to be in a strict agreement to accept (or reject) an argument, in order to collectively accept (or reject) this argument, provided that all other voters are undecided. In details, in the first of a two-stage procedure, the sceptical operator chooses the label `in` (respectively, `out`) as a collective label for an argument when the label `in` (respectively, `out`) is chosen by all individuals. Otherwise, an argument is labeled `undec`. In the first of the two-stage and three-stage procedures in credulous and super credulous operators, the operator chooses the label `in` (respectively `out`) as a collective label for an argument if the label `in` (respectively, `out`) is chosen by some individuals, and all other individuals were undecided about this argument. The purpose of the second and third stages of the three operators is to restore consistency in the collective outcome. The three operators guarantee a compatible outcome, that is, the outcome does not go against the judgment of any individual (refer to the appendix for the formal definitions of the AWPR and SSCOs operators).

The result of using the two types of operators correspond to the two options discussed above. AWPR supports the idea that a label is chosen if it is submitted by the majority regardless what the minority think, while the Sceptical and (Super) Credulous operators (SSCOs) support the idea that minority's opinion should not be completely ignored.

Analyzing these two types of operators using Arrow's principle-based approach does not provide a clear advantage of one over the other. For example, the following two postulates:

> **Collective Rationality** (Awad et al. 2015): the output of the aggregation should be a *complete* labeling.
>
> **Compatibility** (Caminada and Pigozzi 2011): the collective label of every argument does not go against any individual's label of that argument, where `in` and `out` are against each others.

are both satisfied by SSCOs but both violated by AWPR. On the other hand, the following two postulates:

> **Unanimity** (Booth et al. 2014): For each argument, every unanimously agreed upon label is chosen as a collective label.
>
> **Independence** (Awad et al. 2015): the collective label of an argument depends only on the votes on that argument.

are both satisfied by AWPR but both violated by SSCOs. Thus, in the absence of specific preferences over these postulates, other contextual factors may promote the social acceptability of one aggregation rule or the other. In the next section, we consider several such contextual factors and tentatively predict their effects, before reporting a test of these predictions, using *randomized control experiments*, which are the golden standard for doing experimental research.

3 CONTEXTUAL FACTORS

First, and as explained in (Caminada and Pigozzi 2011), one of the main advantages of SSCOs is that their (compatible) outcomes can be defended by every individual who took part in the decision. That is, since the collective evaluation chosen by SSCOs is compatbile with each individual's evaluation in the sense that there is no collectively rejected argument that some individuals think should be accepted, and there is no collectively accepted argument that some individuals think should be rejected, then every individual will feel comfortable defending the collective evaluation in public afterwards. If laypeople perceive this advantage, then experimental manipulations that stress the need for everyone involved to defend the outcome should increase preference for outcomes produced by SSCOs against those produced by AWPR.

Various such manipulations (with regards to what the voters are expected to do afterwards) can be imagined, of which we will test three, from weakest to strongest: a simple reminder of the consequences of the vote; a statement

informing participants that each voter is expected to support and defend the group's decision; and a statement informing participants that should the decision of the group prove a mistake, everyone in the group would share responsibility. We deem this last manipulation the strongest because even if group members accept to defend a decision that passed against their vote, they may not be willing to share the blame in case this decision is mistaken. Indeed Ronnegard (Rönnegard 2015) argued that the attribution of moral responsibility to all members of a committee is legitimate when the decision is taken through unanimous voting, while it is not necessarily the case otherwise.

Second, the difference between AWPR and SSCOs and the principles they use is mostly about whether to ignore the opinion of the minority or not. Accordingly, people may be more comfortable with AWPR outcomes when the minority is very small. To test this hypothesis, we will experimentally manipulate the size of the minority, which will either be small (9 votes to 1) or large (6 votes to 4).

Third, laypeople show a preference for more conservative outcomes when these outcomes may involve the infliction of personal harm, that is, when an individual will incur a cost or a punishment as a consequence of the decision (Bonnefon 2010). Once more, because SSCOs are more conservative than AWPR, then decisions that imply to inflict personal harm should shift preferences towards SSCOs outcomes. To test this hypothesis, we will present participants with decisions that do or do not imply to inflict personal harm.

4 METHOD

A valid comparison of two aggregation rules would require presenting subjects with the detailed explanation of the aggregation rules, and confirming that subjects fully understand the possible outcomes of each rule given any inputs. This is problematic in user-studies regarding formal entailment (like this study) since general (logical) principles are hard to explain to non-logicians.

An alternative approach is to avoid providing technical details and use examples instead, as used in studies of this form (Oaksford and Hahn 2004; Rahwan et al. 2010; Stenning and Cox 1995). In our case, that would be presenting subjects with examples, in which voting profiles and their potential outcomes (given each rule) are provided rather than the explanation of the two rules. However, using this approach, in order to make any claim about how two aggregation rules compare with respect to their favorability by people, a systematic comparison of the two rules is required in exhaustive manner by considering all possible types of inputs e.g. different argument graphs, and different labeling profiles. This would naturally increase the number of factors (to account for these variations), and would exponentially increase the number of conditions and number of needed subjects.

As such, in this study, we only consider a subset of these variations. While these variations are not enough to make claims about how the two aggregation rules compare, they are enough to draw conclusions about the favorability of the principles employed by these rules. We will refer to these principles as the *plurality principle*, that is a label that uniquely has the most votes shall be chosen as a collective label, a principle employed by AWPR, and the *compatibility principle*, that is a label is not collectively accepted (resp. rejected) if it is rejected (resp. accepted) by at least one voter, a principle employed by SSCOs.

A total of 400 participants, all US residents, were recruited from the Crowdflower online platform between March 23rd and May 17th, 2015, and were compensated $0.25 each. Each participant read six vignettes that all featured a committee trying to make a collective decision about a conclusion A (see below for an example). Relevant arguments were listed in each vignette, either two in the case of simple argumentation graphs, or four in the case of complex argumentation graphs (this variable did not impact results, and will not be discussed further). The vignettes displayed

the final collective outcome on each of these arguments, and the task of the participants was to indicate whether, on the basis of these votes, the committee should accept the conclusion A (the AWPR outcome) or declare itself undecided (the SSCOs outcome). Finally, vignettes explained that were the group to declare itself undecided, the decision would be deferred to another body. Full examples are provided in Appendix C.

Of the six vignettes that participants read, three featured a conclusion A that implied to inflict personal harm upon an individual, and three featured a conclusion A that did not imply such a consequence. For example, one vignette featured the conclusion that a football player should be banned for three games (personal harm), whereas one vignette featured the conclusion that the government should stock up on anti-flu drugs (no personal harm).

Participants were randomly assigned to one of eight groups of a 2×4 between-participant design, manipulating the vote ratio and the framing of the decision (i.e., each participant was assigned to one and only one of the eight groups, and each group was presented with six vignettes. The six vignettes per group share the same vote ratio and the same framing of the decision). In one vote ratio condition, all arguments supporting A as well as all arguments rejecting counterarguments to A received 6 votes against 4. In the other vote ratio condition, all arguments supporting A as well as all arguments rejecting counterarguments to A received 9 votes against 1. Finally, the framing of the decision came in four treatments. In the *Baseline* treatment, no information was provided in addition to the above. In the *Reminder* treatment, one sentence reminded participants of the consequences of accepting A (note that this reminder did not contain any new information). In the *Defense* treatment, one sentence explained that each member of the committee was expected to support and defend the committee's decision. In the *Responsibility* treatment, one sentence explained that if the decision of the committee turned out to be wrong, each member would share the responsibility of this mistake.

Table 1. An example of a vignette (including a *no personal harm* scenario) shown to participants in the following condition: *Baseline* treatment, and vote ratio is 6:4. The argumentation graph representing the arguments A, B, and C is similar to the one in Figure 1 (a), but this graph was not shown to participants.

ID: 016

A governmental committee of 10 members gathered to make a collective decision about whether the government should stock up on medicines against the Mexican flu or not. Consider the following main argument:

- *A: The expert virologist says that Mexican flu is a threatening pandemic. Therefore, the government should stock up on medicines against the Mexican flu.*

In considering argument A, the following need to be taken into account:

- *B: This expert has a financial interest in companies making the medicines, according to some journalists. Therefore, his advice cannot be relied upon.*
- *C: This expert does not have a financial interest in the companies making the medicines, according to his employer. Therefore, his advice can be relied upon.*

In the table below, you can see how many members accept ("Yes") or reject ("No") each of the *two* arguments B and C.

No. of Votes	**B**	**C**
6	No	Yes
4	Yes	No

Your job is to determine how they will aggregate their votes. Note that we are not asking about your own opinion on the proposed arguments. Rather, we need to know what you think is appropriate for the group to decide given all the provided information above. The options are:

(1) The group conclude that *the government should stock up on medicines against the Mexican flu.*
(2) The group is undecided about whether *the government should stock up on medicines against the Mexican flu or not.*

In case the committee is undecided, the decision will be postponed for further investigation.

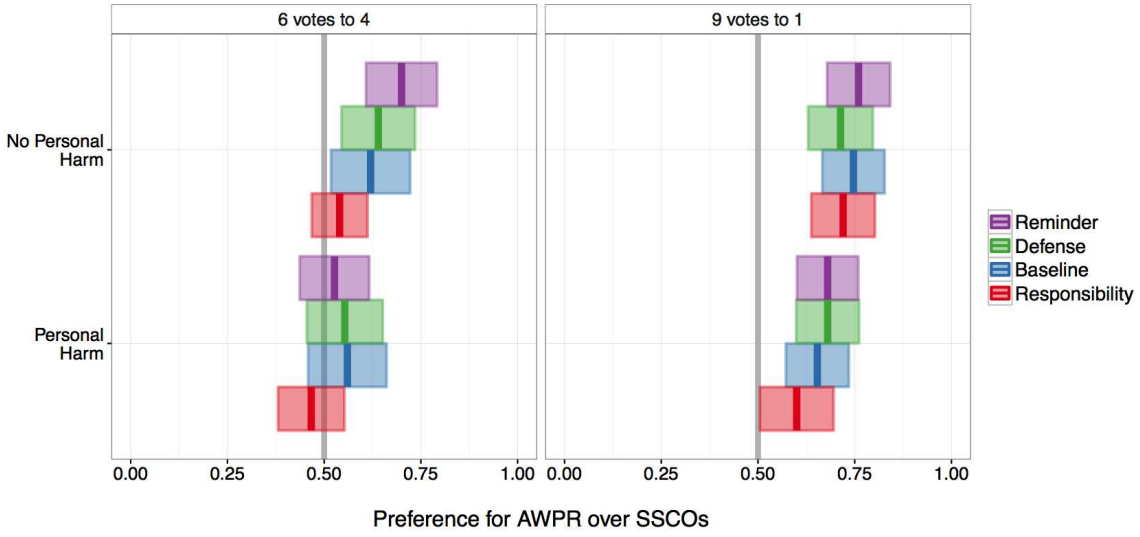


Fig. 3. Preference for AWPR outcomes as a function of decision framing (color-coded), vote ratio, and whether the decision implied to inflict personal harm upon an individual. Boxes shows the average proportion of decisions following AWPR, as well as the 95%-confidence interval around this proportion.

5 RESULTS

Figure 3 displays the average proportion of responses favoring *plurality principle*, as a function of the decision framing (color-coded), the vote ratio, and whether the decision implied to inflict personal harm. The width of each box in Figure 3 indicates the 95%-confidence interval for these proportions. A box that overlaps with the grey line indicates that participants in this condition did not show a significant preference for either principle.

As suggested by Figure 3, participants showed an overall preference for AWPR outcomes. In total, 64% of responses were in line with AWPR outcomes, and this proportion was significantly greater than chance in 11 out of 16 experimental conditions—SSCOs outcomes, on the other hand, were never significantly preferred in any experimental condition.

The preference for AWPR outcomes, though, was qualified by contextual factors. The visual exploration of Figure 3 suggests three main results. First, it seems that the framing of the decision did not significantly affect participants, since the colored boxes are more or less aligned with each other within the four blocks displayed in Figure 3. Second, it appears that participants were less willing to endorse AWPR outcomes in the presence of a large minority, since most of the boxes cross the indifference line in the 6 to 4 condition, whereas none of the boxes cross the indifference line in the 9 to 1 condition. Finally, Figure 3 suggests that participants were less willing to endorse AWPR outcomes when the decision implied to inflict personal harm, given the leftward shift of the boxes in the personal harm condition.

These results were confirmed by a mixed-design analysis of variance in which the dependent variable was the proportion of AWPR responses, the between-participant predictors were the vote ratio and the framing condition, and the within-participant predictor was whether the decision implied to inflict personal harm upon an individual. The results of this analysis are displayed in Table 2. As expected, the analysis detected a statistically significant effect of the vote ratio (64% of AWPR outcomes in the 9 to 1 condition, vs. 58% in the 6 to 4 condition), a statistically significant effect

Table 2. Results of the mixed-design ANOVA analyzing the effect of decision framing, vote ratio, and infliction of personal harm on the preference for AWPR outcomes. The table shows the degree of freedom (DF), the sum of squares (Sum Sq), the mean of squares ($\texttt{Mean Sq} = \frac{\texttt{Sum Sq}}{DF}$), the F test statistics ($F = \frac{\texttt{Mean Sq}}{\texttt{Mean Sq (Residuals)}}$, and the p-value ($P(> F)$). The p-value in the second and the fifth line indicates a significant effect of the vote ratio and infliction of personal harm, respectively.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Framing	3	7.3	2.4	2.0	.11
Ratio	1	25.2	25.2	20.5	< .001
Framing:Ratio	3	0.9	0.3	0.2	.86
Residuals	392	481.3	1.2		
Harm	1	14.6	14.6	25.0	<.001
Framing:Harm	3	1.1	0.4	0.6	.59
Ratio:Harm	1	0.1	0.1	0.2	.64
Framing:Ratio:Harm	3	1.5	0.5	0.9	.45
Residuals	392	228.6	0.6		

of the infliction of personal harm (59% of AWPR outcomes if personal harm, vs. 68% if not), but no effect of the framing of the decision. Post-hoc comparisons using a Bonferroni correction indicated that none of the treatments differed significantly from the Baseline treatment (all $p > .31$). Bonferroni correction (Dunn 1961) is a multiple-comparison method used to adjust for the incorrect rejection of null hypothesis (Type I error) in the case of several simultaneously tested hypotheses. Given these results and the statistical power of our analysis (85% power to detect any effect beyond small, i.e., $f > .25$) we can conclude that even if we had failed to capture a true effect of the framing variable, this effect would likely be small and inconsequential.

6 RELATED WORK

The study of aggregation and voting dates back to the 18th century. Since then, different approaches have been employed in order to decide which method is more appropriate to use. The early years of classical voting witnessed an example-based approach adopted by Borda and Condorcet who separately used examples to show the pitfalls of a voting rule, when compared to their own distinct alternative rules (McLean et al. 1995). Years later, Kenneth Arrow (Arrow 1951; Arrow et al. 2002) adopted a more systematic, principle-based approach by evaluating aggregation rules on the base of satisfying seemingly plausible, "fairness" postulates. Arrow showed the impossibility of satisfying a small set of these postulates together by any aggregation method. This established the superiority of some rules over others, based on subsets of postulates. Thus, in a given scenario, once the desirable postulates are identified, choosing an appropriate rule becomes a systematic task. However, in reality, identifying the desirable postulates in a specific scenario can be subjective, and dependent on complex factors. Aiming to characterize these complex factors, some studies adopted an experiment-based approach (Bassi 2015; Fiorina and Plott 1978; Forsythe et al. 1996; Guarnaschelli et al. 2000; Van der Straeten et al. 2010). These studies established the desirability of some aggregation rules given some assumed conditions from a cognitive perspective.

In the last two decades, aggregation was studied in different new settings including Judgment Aggregation (JA) (Grossi and Pigozzi 2014; List 2012; List and Polak 2010; List and Puppe 2009), non-binary JA (Dokow and Holzman 2010a,b), belief merging (Lin and Mendelzon 1999), labeling aggregation (Awad et al. 2015; Booth et al. 2014; Caminada and Pigozzi 2011; Rahwan and Tohmé 2010), and aggregation of annotated linguistic resources (Endriss and Fernández

2013). While these settings are believed to be closely related, the connection among them and between each of them and the classical voting problem (known as preference aggregation) is not yet fully characterized.

7 DISCUSSION

Formal models of reasoning, argumentation, and decision making can be assessed by intuitions, hypothetical examples and simulations—but also by collecting experimental data in the manner of cognitive psychologists, behavioral economists, and experimental philosophers. Indeed, there is a growing interest in the artificial intelligence community for assessing the cognitive, psychological plausibility of formal models of reasoning (Amgoud et al. 2005; Benferhat et al. 2005; Bonnefon et al. 2008; Dubois et al. 2008; Rahwan et al. 2010). Here, we contribute to this tradition by experimentally identifying the contexts in which human participants would display a preference for SSCOs (aggregation rules that aim at producing compatible outcomes) versus AWPR (the counterpart to the plurality rule in the domain of argumentation).

Our results suggest that outcomes of AWPR were generally more preferred, except in situations where (1) the decision implied to inflict personal harm to an individual, and (2) the vote would pass by a narrow margin. The presence of each of these two factors decreases the preference for AWPR outcomes, and their joint presence leads people to hesitate between AWPR and SSCOs outcomes. However, and interestingly given previous arguments for using SSCOs, (3) we did not observe an increased preference for SSCOs outcomes in situations where all committee members were to defend or take responsibility for the committee's collective decision, independently of their own vote. We now discuss in turn these three findings.

The fact that participants were less likely to endorse the *plurality principle* employed by AWPR for decisions that would result in the infliction of harm to an individual is consistent with previous results showing that people prefer more conservative voting rules in such circumstances (Bonnefon 2007, 2010). This effect may reflect a concern with avoiding costly, unfair false alarms (List 2006), or the emotional saliency of an identified victim (Kogut and Ritov 2006), but its psychological underpinnings are outside the scope of this article. As a consequence of this finding, though, it may be useful to search for novel voting rules which people deem desirable when they have to deliberate on a personal punishment. Indeed, while we observed that the desirability of AWPR outcomes decreased in these circumstances, we did not observe a full switch to a preference for SSCOs outcomes.

We also observe that participants were less comfortable with the *plurality principle* when the vote would pass by a narrow margin (6 to 4 vs. 9 to 1 in our experiment). This result is not surprising: Given that the *plurality principle* effectively ignores the opinion of the minority, it makes sense that people are more comfortable with it when the minority that is ignored is small. The applied consequences of this finding are limited, though. Indeed, shifting to SSCOs when the vote is expected to be a close call essentially amounts to deciding in advance that no decision will be reached.

More importantly, we find no evidence for one purportedly central advantage of the *compatibility principle* employed by SSCOs, that is, their greater desirability when all voters are held accountable (or even responsible) for the committee's decision, independently of their own vote. Participants appeared willing to vote according to AWPR outcomes even in these circumstances. One explanation for this finding is that the *plurality principle* is natural enough to appear justified in most situations. Another possibility is that participants weighted against each other the inconvenience for some to defend a decision they opposed, to the inconvenience for all to defend an indecision nobody voted for. Indeed, if D_o is the inconvenience of defending a decision that one opposed, and D_u is the inconvenience of defending an undecided verdict when one was not undecided, then it might be rational to prefer AWPR when $D_u > \frac{m}{N} \times D_o$, where N is the

total number of voters and m is the number of voters in the minority. For example, in the case of a 6 to 4 vote, it might still be rational (at least from a utilitarian perspective) to apply AWPR when the inconvenience of defending indecision is at least four tenth of the inconvenience of defending the position that one voted against.

In sum, we report the first experimental investigation of the contextual factors that may lead human voters to lean toward SSCOs or AWPR aggregation procedures. Although we observed a general preference for outcomes produced by AWPR, this preference was moderated by contextual factors. We found null evidence against one important prior claim (SSCOs are good when everyone is expected to defend the collective decision), positive evidence for a vote ratio effect (people like AWPR better when the minority is small), and positive evidence for a moral effect (people like AWPR less when they deliberate about inflicting personal harm to an individual).

These results provide clear suggestions regarding what aggregation rules are more favorable to use in some contexts, while at the same time they identify contexts in which the favorability of an aggregation rule over another is still open. This opens the door for further work to explore the favorability of new rules in such contexts, on the road towards eventually offering a comprehensive mapping from a given conflicting-domain context to the appropriate aggregation rule to use in that context.

It is important to note that our study compares the two rules using only some variations of argumentation graphs and labeling profiles. Further, the stimuli used probe the favorability of the outcomes of the aggregation rules (rather than the rules themselves). These outcomes can coincide with outcomes by other rules not studied here. As such, the results above are better interpreted on the level of the principles employed by these rules rather than on the level of the rules themselves. In order to make stronger claims regarding the two rules, the consideration of an exhasutive comparison is required, which can be a topic for future work. Future research will also identify the psychological mechanisms underlying these preferences, but also the voting rules that people may approve of in situations (e.g., personal harm plus narrow margin) where they seem unsatisfied with AWPR and SSCOs both.

ACKNOWLEDGMENTS

Awad is grateful for the fund provided by Masdar Institute to run the experiments. Support through the ANR-Labex IAST is gratefully acknowledged by Bonnefon.

APPENDIX

A LABELING AGGREGATION METHODS

For this appendix, we write $\texttt{in}(L)$, $\texttt{out}(L)$, and $\texttt{undec}(L)$ for the set of arguments that are labeled in, out, and undec respectively by labeling L. A labeling L can be represented as $L = (\texttt{in}(L),\texttt{out}(L),\texttt{undec}(L))$. Equivalently, we also denote a labeling L as: $L = \{(A, l)|L(A) = l \text{ for all } A \in \mathcal{A}, l \in \{\texttt{in}, \texttt{out}, \texttt{undec}\}\}$.

The problem of labeling aggregation can be formulated as a set of individuals that collectively decide how an argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ must be labelled.

Definition A.1 (Labeling aggregation problem (Awad et al. 2015)). Let $Ag = \{1, \ldots, n\}$ be a finite non-empty set of agents, and $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. A labeling aggregation problem is a pair $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$.

Each individual $i \in Ag$ has a labeling L_i which expresses the evaluation of $\mathcal{AF}$ by this individual. A labeling profile is an $|Ag|$-tuple of labelings.

Definition A.2 (Labeling profile (Awad et al. 2015)). Let $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$ be a labeling aggregation problem. We use $\mathcal{L} = (L_1, \ldots, L_n) \in \mathbf{L}(\mathcal{AF})^{|Ag|}$ to denote a labeling profile, where $\mathbf{L}(\mathcal{AF})$ is the class of labelings of $\mathcal{AF}$. Additionally, we use $\mathcal{L}(a)$ to denote the labeling profile (i.e. an $|Ag|$-tuple) of an argument $a \in \mathcal{A}$ i.e. $\mathcal{L}(a) = (L_1(a), \ldots, L_n(a))$.

The aggregation of individuals' labelings can be defined as a partial function.[2]

Definition A.3 (Aggregation function (Awad et al. 2015)). Let $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$ be a labeling aggregation problem. An aggregation function for $\mathcal{LAP}$ is a function $F : \mathbf{L}(\mathcal{AF})^n \to \mathbf{L}(\mathcal{AF})$.

For each $a \in \mathcal{A}$, $[F(\mathcal{L})](a)$ denotes the collective label assigned to a, if F is defined for $\mathcal{L} = (L_1, \ldots, L_n)$.

A.1 The Argument-Wise Plurality Rule

The *Argument-Wise Plurality Rule (AWPR) M* was proposed in (Awad et al. 2015; Rahwan and Tohmé 2010). Intuitively, for each argument, it selects the label that appears most frequently in the individual labelings.

Definition A.4 (Argument-Wise Plurality Rule (AWPR) (Awad et al. 2015)). Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. Given any argument $a \in \mathcal{A}$ and any profile $\mathcal{L} = (L_1, \ldots, L_n)$, it holds that $[M(\mathcal{L})](a) = l_a \in \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$ iff

$$|\{i : L_i(a) = l_a\}| > \max_{l'_a \neq l_a} |\{i : L_i(a) = l'_a\}|$$

Note that M is defined for all profiles that cause no ties, i.e. $M(\mathcal{L})$ is defined iff there does not exist any argument $a \in \mathcal{A}$ for which we have at least two labels l_a and l'_a with $l_a \neq l'_a$ and

$$|\{i : L_i(a) = l_a\}| = |\{i : L_i(a) = l'_a\}| = \max_{l} |\{i : L_i(a) = l\}|$$

A.2 Sceptical and (Super) Credulous Operators (SSCOs)

The three aggregation operators, namely the *Sceptical*, the *Credulous* and the *Super Credulous*, were defined in (Caminada and Pigozzi 2011). In their work, a labeling profile is represented as a set, since the number of votes for each label is irrelevant (as long as it is not zero) for these operators. For convenience, we will assume here that the labeling profile is a tuple (as in Def. A.2).

Crucial to the definitions of the three operators are the notions of *less or equally committed* and *compatible*. A labeling L_1 is said to be *less or equally committed* than another labeling L_2 if and only if every argument that is labeled $\mathtt{in}$ by L_1 is also labeled $\mathtt{in}$ by L_2 and every argument that is labeled $\mathtt{out}$ by L_1 is also labeled $\mathtt{out}$ by L_2.

Definition A.5 (Less or equally committed $\sqsubseteq$ (Caminada and Pigozzi 2011)). Let L_1 and L_2 be two labelings of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. L_1 is less or equally committed as L_2 ($L_1 \sqsubseteq L_2$) iff $\mathtt{in}(L_1) \subseteq \mathtt{in}(L_2)$ and $\mathtt{out}(L_1) \subseteq \mathtt{out}(L_2)$.

Two labelings L_1 and L_2 are said to be *compatible* with each other if and only if for every argument, there is no $\mathtt{in} - \mathtt{out}$ conflict between the two. In other words, every argument that is labeled $\mathtt{in}$ by L_1 is not labeled $\mathtt{out}$ by L_2 and every argument that is labeled $\mathtt{out}$ by L_1 is not labeled $\mathtt{in}$ by L_2.

Definition A.6 (Compatible labelings $\approx$ (Caminada and Pigozzi 2011)). Let L_1 and L_2 be two labelings of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We say that L_1 is compatible with L_2 ($L_1 \approx L_2$) iff $\mathtt{in}(L_1) \cap \mathtt{out}(L_2) = \emptyset$ and $\mathtt{out}(L_1) \cap \mathtt{in}(L_2) = \emptyset$

[2] We state that the function is partial to allow for cases in which collective judgment may be undefined (e.g. when there is a tie in voting).

The following two definitions are used in the definition of the operators:

Definition A.7 (Initial operators $\sqcap$, $\sqcup$ (Caminada and Pigozzi 2011)). The sceptical initial $\sqcap$ and credulous initial $\sqcup$ operators are labeling aggregation operators defined as the following:

- $\sqcap((L_1,\ldots,L_n)) = \{(A,\texttt{in})|\forall i \in Ag : L_i(A) = \texttt{in}\} \cup \{(A,\texttt{out})|\forall i \in Ag : L_i(A) = \texttt{out}\} \cup \{(A,\texttt{undec})|\exists i \in Ag : L_i(A) \neq \texttt{in} \wedge \exists j \in Ag : L_j(A) \neq \texttt{out}\}$
- $\sqcup((L_1,\ldots,L_n)) = \{(A,\texttt{in})|\exists i \in Ag : L_i(A) = \texttt{in} \wedge \neg\exists j \in Ag : L_j(A) = \texttt{out}\} \cup \{(A,\texttt{out})|\exists i \in Ag : L_i(A) = \texttt{out} \wedge \neg\exists j \in Ag : L_j(A) = \texttt{in}\} \cup \{(A,\texttt{undec})|\forall i \in Ag : L_i(A) = \texttt{undec} \vee (\exists j \in Ag : L_j(A) = \texttt{in} \wedge \exists k \in Ag : L_k(A) = \texttt{out})\}$

Definition A.8 (Down-admissible $\downarrow$ and up-complete $\uparrow$ labelings (Caminada and Pigozzi 2011)). Let L be a labeling of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. The down-admissible labeling of L, denoted as $L\downarrow$, is the biggest element of the set of all admissible labelings that are less or equally committed than L. The up-complete labeling of L, denoted as $L\uparrow$, is the smallest element of the set of all complete labelings that are bigger or equally committed than L.

Where *complete* labelings are labelings that satify the three conditions in Def. 2.3, *admissible* labelings are labelings that satisfy the first two conditions of Def. 2.3, and *biggest* and *smallest* are defined as with respect to Def. A.5. Now, we provide the definition of the three operators:[3]

Definition A.9 (SSCOs: $so_{\mathcal{AF}}$, $co_{\mathcal{AF}}$ and $sco_{\mathcal{AF}}$ (Caminada and Pigozzi 2011)). Given an argumentation framework $\mathcal{AF}$, the sceptical $so_{\mathcal{AF}}$, the credulous $co_{\mathcal{AF}}$ and the super credulous $sco_{\mathcal{AF}}$ operators are labeling aggregation operators defined as the following:

- $so_{\mathcal{AF}}((L_1,\ldots,L_n)) = (\sqcap((L_1,\ldots,L_n)))\downarrow$.
- $co_{\mathcal{AF}}((L_1,\ldots,L_n)) = (\sqcup((L_1,\ldots,L_n)))\downarrow$.
- $sco_{\mathcal{AF}}((L_1,\ldots,L_n)) = ((\sqcup((L_1,\ldots,L_n)))\downarrow)\uparrow$.

Note that the super credulous operator was introduced since the credulous operator can fail to produce a complete collective labeling. This is not the case for sceptical operator, which always produces complete collective labeling.

A.3 Postulates

Inspired by Arrow's principle-based approach, many postulates were defined in the context of argumentation. Most of these postulates are similar to the ones proposed in Judgment Aggregation. We provide here the formal definition of the four postulates that were mentioned in the introduction. *Collective rationality* requires that any possible outcome of the aggregation rule has to be a complete labeling.

> **Collective Rationality** (Awad et al. 2015) For all profiles $\mathcal{L}$ such that $F(\mathcal{L})$ is defined, $F(\mathcal{L}) \in Comp(\mathcal{AF})$.

where $Comp(\mathcal{AF})$ is the set of all possible complete labelings for the argumentation framrework $\mathcal{AF}$. *Compatibility* requires that the collective label of each argument is compatible (w.r.t Def. A.6) with the label by every individual for that argument.

> **Compatibility** (Caminada and Pigozzi 2011) For all $i \in Ag$ and $a \in \mathcal{A}$ we have $[F_{\mathcal{AF}}(\mathcal{L})](a) \approx L_i(a)$.

Unanimity requires that for each argument, every unanimously agreed upon label is chosen as a collective label.

[3]For a better understanding of the three operators, the reader is encouraged to see the clarifying examples of the three operators in (Caminada and Pigozzi 2011).

Unanimity (Booth et al. 2014) For each $a \in \mathcal{A}$, if there is some $\mathtt{x} \in \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$ such that $L_i(a) = \mathtt{x}$ for all $i \in Ag$ then $[F_{\mathcal{AF}}(\mathcal{L})](a) = \mathtt{x}$.

Finally, *independence* requires that the collective label of an argument depends only on the votes on that argument.

Independence (Awad et al. 2015) For any two profiles $\mathcal{L} = (L_1, \ldots, L_n)$, $\mathcal{L}' = (L'_1, \ldots, L'_n)$ such that $F(\mathcal{L})$ and $F(\mathcal{L}')$ are defined, and for all $a \in \mathcal{A}$, if $L_i(a) = L'_i(a)$ for all $i \in Ag$, then $[F(\mathcal{L})](a) = [F(\mathcal{L}')](a)$.

B ARGUMENT SETS

Following, are the six stories that were used in the experiment. The argumentation structure representing the first three is shown in Figures 4 and the one representing the other three is shown in Figure 5.

B.1 Simple AF (shown in Figure 4)

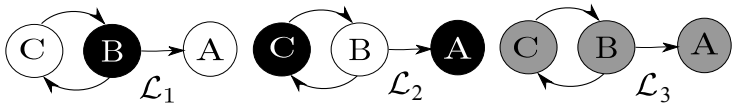


Fig. 4. A simple argumentation graph with the three possible *complete labelings*. Nodes are arguments, and arcs are the defeat relations. The color white refers to in, the color black refers to out, and the color gray refers to undec.

(1) Argument set 1 - Argument from Knowledge

Context: A governmental committee of 10 members gathered to make a collective decision about whether the government should stock up on medicines against the Mexican flu or not.

- A: The expert virologist says that Mexican flu is a threatening pandemic. Therefore, the government should stock up on medicines against the Mexican flu.
- B: This expert has a financial interest in companies making the medicines, according to some journalists. Therefore, his advice cannot be relied upon.
- C: This expert does not have a financial interest in the companies making the medicines, according to his employer. Therefore, his advice can be relied upon.

In case the committee is undecided, the decision will be postponed for further investigation.

(2) Argument set 2 - Argument from Analogy, Classification and Precedent

Context: A school committee of 10 teachers gathered to make a collective decision about whether the school should have a uniform or not.

- A: After the Central Public School forced students to have a uniform, the attendance of students increased. Therefore, we should have a uniform in our school.
- B: There can be other factors that contributed to the effect of uniform on attendance. Their case might be different from ours.
- C: We share the same entry standards with the Central Public School and our student's families have similar socio-economic status to theirs.[4]

[4]One might note that, in reality, argument C might not defeat B. We only noticed that this can be the case after the experiment was over. However, upon removing this example and redoing the analysis, no major change was found in the results.

In case the committee is undecided, the decision will be deferred to the school principal who will form another committee.

(3) Argument set 3 - Argument from Knowledge

Context: A hiring committee of 10 members gathered to make a collective decision about whether a specific candidate is worthy of a good offer or not.

- A: The candidate's former adviser provided a strong recommendation letter. Therefore, this candidate is worthy of a good offer.
- B: It is in the adviser's interest that her former student gets a good position. Therefore, she is probably over-selling him.
- C: The adviser knows she can lose her credibility if the candidate is not as good as she claims.

In case the committee is undecided, the decision will be postponed to get further reference letters.

B.2 Non-Simple AF (shown in Figure 5)

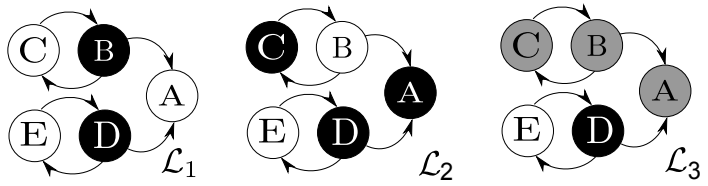


Fig. 5. A non-simple argumentation graph with three possible *complete labelings*. Nodes are arguments, and arcs are the defeat relations. The color white refers to in, the color black refers to out, and the color gray refers to undec.

(1) Argument set 1 - Argument from Knowledge

Context: A committee of 10 jury members gathered to make a collective decision about whether there is evidence against Stephen or not.

- A: The witness saw Stephen in the parking area next to the crime scene. Therefore, there is evidence against Stephen.
- B: It was dark then. Therefore, the witness probably mistook someone else for Stephen.
- C: The parking area is well lit. Therefore, the witness could clearly identify Stephen.
- D: The witness hates Stephen. Therefore, the witness is biased.
- E: The witness did not know Stephen well. Therefore, the witness is unbiased.

In case the committee is undecided, the decision will be deferred to the judge who will form another committee.

(2) Argument set 2 - Argument from Knowledge

Context: A group of 10 employees were assigned the task of making a collective decision about whether the company's next summer excursion should be to Niagara Falls or not.

- A: The travel agent recommended visiting Niagara Falls. Therefore, the next summer excursion should be to Niagara Falls.
- B: The travel agent has never been to Niagara Falls. Therefore, we cannot trust his recommendation.
- C: The travel agent has organized many trips to Niagara Falls before. Therefore, we can trust his recommendation.

- D: The travel agent recommended visiting a place with natural attractions citing Niagara Falls as an example. He did not specifically recommend Niagara Falls.
- E: The travel agent specifically recommended visiting Niagara Falls citing its natural attractions as the main reason.

In case the committee is undecided, the decision will be deferred to the senior management who will form another committee.

(3) Argument set 3 - Argument from Analogy, Classification and Precedent

Context: A referees committee of 10 members gathered to make a collective decision about whether the footballer Marconi should be banned for three matches or not.

- A: Player Marconi criticized the referee via his official Twitter account. In a previous incident, the footballer Borello was banned for three matches for publicly criticizing the referee. Therefore, Marconi should be banned for three matches.
- B: Borello criticized the referee in a press conference. That was a different case.
- C: Both cases are similar in what it matters.
- D: In another previous incident, the footballer Zotti criticized a referee and got away with it. That was a similar case.
- E: Zotti's criticism was less direct than the criticism by Marconi and Borello.

In case the committee is undecided, the decision will be postponed for further investigation.

C VIGNETTES

Here are some concrete examples of vignettes using the argument sets above.

C.1 Simple AF, Case: ratio 9 : 1, *reminder*, Scenario: Uniform

ID: 229
A school committee of 10 teachers gathered to make a collective decision about whether the school should have a uniform or not. Consider the following main argument:

- *A: After the Central Public School forced students to have a uniform, the attendance of students increased. Therefore, we should have a uniform in our school.*

In considering argument A, the following arguments need to be taken into account:

- *B: There can be other factors that contributed to the effect of uniform on attendance. Their case might be different from ours.*
- *C: We share the same entry standards with the Central Public School and our student's families have similar socio-economic status to theirs.*

In the table below, you can see how many members accept ("Yes") or reject ("No") each of the *two* arguments B and C.

No. of Votes	B	C
9	No	Yes
1	Yes	No

Your job is to determine how they will aggregate their votes. Note that we are not asking about your own opinion on the proposed arguments. Rather, we need to know what you think is appropriate for the group to decide given all the provided information above. The options are:

(1) The group conclude that the school should have a uniform.
[This will add extra costs on the students' parents' side]
(2) The group is undecided about whether the school should have a uniform or not.
[In this case, the decision will be deferred to the school principal who will form another committee]

C.2 Complex AF, Case: ratio 9 : 1, *defending*, Scenario: Marconi

ID: 169
A referees committee of 10 members gathered to make a collective decision about whether the footballer Marconi should be banned for three matches or not. Consider the following main argument:

- *A: Player Marconi criticized the referee via his official Twitter account. In a previous incident, the footballer Borello was banned for three matches for publicly criticizing the referee. Therefore, Marconi should be banned for three matches.*

In considering argument A, the following arguments need to be taken into account:

- *B: Borello criticized the referee in a press conference. That was a different case.*
- *C: Both cases are similar in what it matters.*
- *D: In another previous incident, the footballer Zotti criticized a referee and got away with it. That was a similar case.*
- *E: Zotti's criticism was less direct than the criticism by Marconi and Borello.*

In the table below, you can see how many members accept ("Yes") or reject ("No") each of the *four* arguments B, C, D, and E.

No. of Votes	B	C	D	E
9	No	Yes	No	Yes
1	Yes	No	Yes	No

Your job is to determine how they will aggregate their votes. Note that we are not asking about your own opinion on the proposed arguments. Rather, we need to know what you think is appropriate for the group to decide given all the provided information above. The options are:

(1) The group conclude that *Marconi should be banned for three matches.*
(2) The group is undecided about whether *Marconi should be banned for three matches or not.*

In case the committee is undecided, the decision will be postponed for further investigation.

Note that once a collective decision is made, each committee member is expected to support and defend it. It may be awkward for a committee member who disagrees with the group conclusion to defend it to others.

C.3 Complex AF, Case: ratio 6 : 4, *responsibility*, Scenario: Excursion

ID: 356
A group of 10 employees were assigned the task of making a collective decision about whether the company's next summer excursion should be to Niagara Falls or not. Consider the following main argument:

- *A: The travel agent recommended visiting Niagara Falls. Therefore, the next summer excursion should be to Niagara Falls.*

In considering argument A, the following arguments need to be taken into account:

- *B: The travel agent has never been to Niagara Falls. Therefore, we cannot trust his recommendation.*
- *C: The travel agent has organized many trips to Niagara Falls before. Therefore, we can trust his recommendation.*
- *D: The travel agent recommended visiting a place with natural attractions citing Niagara Falls as an example. He did not specifically recommend Niagara Falls.*
- *E: The travel agent specifically recommended visiting Niagara Falls citing its natural attractions as the main reason.*

In the table below, you can see how many members accept ("Yes") or reject ("No") each of the *four* arguments B, C, D, and E.

No. of Votes	B	C	D	E
6	No	Yes	No	Yes
4	Yes	No	Yes	No

Your job is to determine how they will aggregate their votes. Note that we are not asking about your own opinion on the proposed arguments. Rather, we need to know what you think is appropriate for the group to decide given all the provided information above. The options are:

(1) The group conclude that the company's next summer excursion should be to Niagara Falls.
[If this decision turned out to be wrong, each member would share the responsibility for this mistake]
(2) The group is undecided about whether the company's next summer excursion should be to Niagara Falls or not.
[In this case, the decision will be deferred to the senior management who will form another committee]

REFERENCES

Leila Amgoud, Jean-François Bonnefon, and Henri Prade. 2005. An argumentation-based approach for multiple criteria decision. *Lecture Notes in Computer Science* 3571 (2005), 269–280.

Gordana Apic, Matthew Betts, and Robert Russell. 2011. Content disputes in Wikipedia reflect geopolitical instability. *PloS one* 6, 6 (2011), e20902.

Kenneth J. Arrow. 1951. *Social choice and individual values*. Wiley, New York NY, USA.

Kenneth J. Arrow, A. K. Sen, and K. Suzumura (Eds.). 2002. *Handbook of Social Choice and Welfare*. Vol. 1. Elsevier Science Publishers (North-Holland).

Edmond Awad, Richard Booth, Fernando Tohme, and Iyad Rahwan. 2015. Judgment Aggregation in Multi-Agent Argumentation. *Journal of Logic and Computation* (2015). DOI: http://dx.doi.org/10.1093/logcom/exv055

Anna Bassi. 2015. Voting systems and strategic manipulation: An experimental study. *Journal of Theoretical Politics* 27, 1 (2015), 58–85.

Salem Benferhat, Jean-François Bonnefon, and Rui da Silva Neves. 2005. An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese* 146 (2005), 53–70. DOI: http://dx.doi.org/10.1007/s11229-005-9069-6

Jean-François Bonnefon. 2007. How do individuals solve the doctrinal paradox in collective decisions? An empirical investigation. *Psychological Science* 18 (2007), 753–755.

Jean-François Bonnefon. 2010. Behavioral evidence for framing effects in the resolution of the doctrinal paradox. *Social choice and welfare* 34, 4 (2010), 631–641.

Jean-François Bonnefon, Didier Dubois, Hélène Fargier, and Sylvie Leblois. 2008. Qualitative heuristics for balancing the pros and cons. *Theory and Decision* 65 (2008), 71–95.

Richard Booth, Edmond Awad, and Iyad Rahwan. 2014. Interval methods for judgment aggregation in argumentation. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR 2014)*. 594–597.

Luc Bovens and Wlodek Rabinowicz. 2006. Democratic answers to complex questions–an epistemic perspective. *Synthese* 150, 1 (2006), 131–153.

Simon Buckingham Shum. 2008. Cohere: Towards Web 2.0 Argumentation. In *Proceedings of the 2nd International Conference on Computational Models of Argument (COMMA)*, Philippe Besnard, Sylvie Doutre, and Anothony Hunter (Eds.). IOS Press, Amsterdam, The Netherlands, 97–108.

Martin Caminada. 2006. On the Issue of Reinstatement in Argumentation. In *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA)*, Michael Fisher, Wiebe van der Hoek, Boris Konev, and Alexei Lisitsa (Eds.). Lecture Notes in Computer Science, Vol. 4160. Springer, 111–123.

Martin Caminada and Dov M. Gabbay. 2009. A Logical Account of Formal Argumentation. *Studia Logica* 93, 2-3 (2009), 109–145.

Martin Caminada and Gabriella Pigozzi. 2011. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems* 22, 1 (2011), 64–102.

Elad Dokow and Ron Holzman. 2010a. Aggregation of binary evaluations with abstentions. *Journal of Economic Theory* 145, 2 (2010), 544–561.

Elad Dokow and Ron Holzman. 2010b. Aggregation of non-binary evaluations. *Advances in Applied Mathematics* 45, 4 (2010), 487–504.

Didier Dubois, Hélène Fargier, and Jean-François Bonnefon. 2008. On the qualitative comparison of decisions having positive and negative features. *Journal of Artificial Intelligence Research* 32 (2008), 385–417.

Phan M. Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77, 2 (1995), 321–358.

Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.

Ulle Endriss and Raquel Fernández. 2013. Collective annotation of linguistic resources: Basic principles and a formal model. In *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence, Delft, The Netherlands, November 7-8, 2013*. Delft University of Technology (TU Delft); under the auspices of the Benelux Association for Artificial Intelligence (BNVKI) and the Dutch Research School for Information and Knowledge Systems (SIKS).

Morris P Fiorina and Charles R Plott. 1978. Committee decisions under majority rule: An experimental study. *American Political Science Review* 72, 02 (1978), 575–598.

Robert Forsythe, Thomas Rietz, Roger Myerson, and Robert Weber. 1996. An experimental study of voting rules and polls in three-candidate elections. *International Journal of Game Theory* 25, 3 (1996), 355–383.

Davide Grossi and Gabriella Pigozzi. 2014. *Judgment Aggregation: A Primer*. Morgan & Claypool.

Serena Guarnaschelli, Richard D McKelvey, and Thomas R Palfrey. 2000. An experimental study of jury decision rules. *American Political Science Review* 94, 02 (2000), 407–423.

Joel M Guttman. 1998. Unanimity and majority rule: the calculus of consent reconsidered*. *European Journal of Political Economy* 14, 2 (1998), 189–207.

Joshua Introne, Robert Laubacher, Gary Olson, and Thomas Malone. 2011. The Climate CoLab: Large scale model-based collaborative planning. In *Collaboration Technologies and Systems (CTS), 2011 International Conference on*. IEEE, 40–47.

Mark Klein and Luca Iandoli. 2008. Supporting Collaborative Deliberation Using a Large-Scale Argumentation System: The MIT Collaboratorium. *Available at SSRN 1099082* 4691-08 (2008).

Tehila Kogut and Ilana Ritov. 2006. The "identified victim" effect: an identified group, or just a single individual? *Journal of Behavioral Decision Making* 18 (2006), 157–167.

Jinxin Lin and Alberto O Mendelzon. 1999. Knowledge base merging by majority. In *Dynamic Worlds*. Springer, 195–218.

Christian List. 2006. The discursive dilemma and public reason. *Ethics* 18 (2006), 362–402.

Christian List. 2012. The theory of judgment aggregation: An introductory review. *Synthese* 187, 1 (2012), 179–207.

Christian List and Ben Polak. 2010. Introduction to judgment aggregation. *Journal of economic theory* 145, 2 (2010), 441–466.

Christian List and Clemens Puppe. 2009. Judgment aggregation: a survey. In *The handbook of rational and social choice*, Paul Anand, Clemens Puppe, and Prasanta Pattanaik (Eds.). Oxford University Press, Oxford, UK.

Seth Marvel, Jon Kleinberg, Robert Kleinberg, and Steven Strogatz. 2011. Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences* 108, 5 (2011), 1771–1776.

Iain McLean, Arnold B Urken, and Fiona Hewitt. 1995. *Classics of social choice.* University of Michigan Press.

Luis Miller and Christoph Vanberg. 2015. Group size and decision rules in legislative bargaining. *European Journal of Political Economy* 37 (2015), 288–302.

Mike Oaksford and Ulrike Hahn. 2004. A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 58, 2 (2004), 75.

Philip Pettit. 2001. Deliberative democracy and the discursive dilemma. *Noûs* 35, s1 (2001), 268–299.

Gabriella Pigozzi. 2006. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese* 152, 2 (2006), 285–298.

Antonio Quesada. 2011. Parallel axiomatizations of majority and unanimity. *Economics Letters* 111, 2 (2011), 151–154.

Iyad Rahwan, Mohammed I Madakkatel, Jean-François Bonnefon, Ruqiyabi N Awan, and Sherief Abdallah. 2010. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science* 34, 8 (2010), 1483–1502.

Iyad Rahwan and Fernando Tohmé. 2010. Collective argument evaluation as judgement aggregation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 417–424.

Iyad Rahwan, Fouad Zablith, and Chris Reed. 2007. Laying the foundations for a world wide argument web. *Artificial intelligence* 171, 10 (2007), 897–921.

David Rönnegard. 2015. *The Fallacy of Corporate Moral Agency.* Springer.

Keith Stenning and Richard Cox. 1995. Attitudes to logical independence: traits in quantifier interpretion. (1995).

Karine Van der Straeten, Jean-François Laslier, Nicolas Sauger, and André Blais. 2010. Strategic, sincere, and heuristic voting under four election rules: an experimental study. *Social Choice and Welfare* 35, 3 (2010), 435–472.