Accountable Approval Sorting

Khaled Belahcene[1], Yann Chevaleyre[2], Christophe Labreuche[3], Nicolas Maudet[4], Vincent Mousseau[1], Wassila Ouerdane[1]
[1] Laboratoire Genie Industriel, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
[2] Université Paris-Dauphine, PSL Research University, CNRS, UMR [7243], LAMSADE, France
[3] Thales Research and Technology, Palaiseau, France
[4] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, France
{khaled.belahcene, vincent.mousseau, wassila.ouerdane}@centralesupelec.fr,
yann.chevaleyre@dauphine.fr, christophe.labreuche@thalesgroup.com, nicolas.maudet@lip6.fr

Abstract

We consider decision situations in which a set of points of view (voters, criteria) are to sort a set of candidates to ordered classes (GOOD / BAD). Candidates are judged GOOD when approved by a sufficient set of points of view; this corresponds to noncompensatory sorting. To be accountable, such approval sorting should provide guarantees about the decision process and decisions concerning specific candidates. We formalize accountability using a feasibility problem expressed as a boolean satisfiability formulation. We illustrate different forms of accountability when a committee decides with approval sorting and study the information that should be disclosed by the committee.

1 Introduction

A committee meets to decide upon the sorting of a number of candidates into two categories (e.g. candidates to accept or not, projects to fund or not). The committee applies a decision process which is public, the outcomes are public as well, however the details of the votes are sensitive and should not be made available. Recently, the issue of the *accountability* of algorithmic decisions has become a primary concern of our society [Doshi-Velez *et al.*, 2017; Wachter *et al.*, 2017]. To what extent can we make the committee accountable of its decisions? In particular, in our setting, a distinctive feature is that the decision may concern several individuals: being accountable for the classification of an individual may not be the same as being accountable for all the classifications. To make things more precise, it is thus useful to distinguish the following situations:

S1: an independent audit agency is commissioned to check that the decisions of the committee indeed comply with the publicly announced decision rule.

S2: a candidate, (supposedly) unsatisfied with the outcome of the process regarding his own classification, challenges the committee and asks for a justification.

Situation S1 is sometimes called *procedural regularity*, see for instance [Kroll *et al.*, 2017], which calls for systems able to prove to oversight authorities that "decisions are made under an announced set of rules consistently applied in each case". A typical way to address situation S1 is to require *transparency* and let the audit agency access all the available information. This suffers from two drawbacks: (i) there are often exceptions making full disclosure of the decision procedure impossible, (ii) the burden of proof lies on the shoulders of the audit agency, which (depending on the model) may be too demanding. Alternatively, we can leave the burden of proof on the committee's side and ask for evidence that the set of classifications is compliant with the decision process. This may be done by exhibiting only part of the information, illustrating that the obtained classification is a *possible* outcome of the sorting process. Since, typically, many other outcomes would also be possible, this could preserve to some extent the privacy of the committee's votes. On the other hand, failing this test would be evidence that the process was biased.

Regarding situation S2, the objective is to justify the classification of the complaining individual, again with minimal disclosure of the committee's votes. In this case, the committee will aim for evidence that the classification of the candidate cannot be otherwise, *as long as a number of other classification outcomes are accepted*. We can think of such decisions as reference cases. Technically, this requires to show the impossibility to rank the candidate in a different category, *i.e.* the decision is *necessary* with respect to the jurisprudence.

More precisely, we shall primarily be concerned with a general sorting model where voters express binary judgments [Laslier and Sanver, 2010], and candidates are sorted as either *good* or *bad* depending on the fact that the coalition of voters supporting this classification is winning or not. An important hypothesis is that the set of winning coalitions has to remain constant for the set of classifications under scrutiny. This can be seen as a requirement for the process to be unbiased. In this setting, the "details of the votes" cover two aspects: (i) the approval of voters at the individual level, (ii) the winning coalitions at the committee level. In this paper we address the following research question:

Can we make the decisions of a committee using approval sorting accountable while preserving as much as possible the details of the votes?

The details of the sorting model are given in Section 2. At the core of our proposal lies a characterization result of the sorting model which avoids explicit reference to winning coalitions, and leads to a SAT encoding (Section 3). In Section 4, we consider the different scenarios discussed in the introduction and show how this formal machinery allows us to provide argument schemes which answer, at least partially, the accountability requirements. Section 5 discusses related work and concludes.

2 Noncompensatory Sorting

We are interested in situations where there is a need to aggregate diverse, potentially conflicting, *points of view* forming a set $\mathcal{N}$ – each $i \in \mathcal{N}$ can be seen as an agent, a voter, or a criterion – into a single *sorting* of some *alternatives* taken in a set $\mathbb{X}$ between two categories, GOOD and BAD, expressed by an *assignment* $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$. Each point of view $i \in \mathcal{N}$ has an opinion on the entire set of alternatives in the form of a complete preorder $\succsim_i$ (*i.e.* $\succsim_i$ is a complete, reflexive and transitive binary relation on $\mathbb{X}$) . This preference may stem from numeric or symbolic performance, as it is often the case in multi-criteria decision aiding, or be intrinsically ordinal, as it is often assumed in social choice contexts. Nevertheless, the aggregation procedure requires that each point of view $i \in \mathcal{N}$ expresses only a binary judgment on each alternative $x \in \mathbb{X}$ which is either approved or not according to i. We shall also consider a subset $\mathbb{X}^\star \subseteq \mathbb{X}$ of alternatives with a *reference* status, with their assignment $\alpha^\star : \mathbb{X}^\star \to \{\text{GOOD}, \text{BAD}\}$ serving as a basis for elaborating justifications.

This abstract description covers several well-documented decision processes, e.g. :

- a multiple criteria sorting problem [Bouyssou *et al.*, 2006] with ordinal preferences (each point of view $i \in \mathcal{N}$ is a *criterion*);
- a committee decision context (each point of view $i \in \mathcal{N}$ is a *voter* and the GOOD category is the set of winners).

Example 1. *We consider a situation with six alternatives* $\mathbb{X} := \{a, b, c, d, e, f\}$*, assessed from five points of view* $\mathcal{N} := \{1, 2, 3, 4, 5\}$ *in the following manner:*

$$a \succ_1 b \succ_1 f \succ_1 e \succ_1 c \succ_1 d$$
$$e \succ_2 b \succ_2 c \succ_2 d \succ_2 a \succ_2 f$$
$$f \succ_3 a \succ_3 b \succ_3 d \succ_3 e \succ_3 c$$
$$d \succ_4 a \succ_4 c \succ_4 e \succ_4 f \succ_4 b$$
$$c \succ_5 e \succ_5 b \succ_5 f \succ_5 d \succ_5 a$$

We recall the definitions of an upset and the upper closure of a subset w.r.t. a binary relation:

Definition 1 (Upset and upper closure)**.** *Let A be a set and $\mathcal{R}$ a binary relation on A. An upset of $(A, \mathcal{R})$ is a subset $B \subseteq A$ such that $\forall a \in A, \forall b \in B, a\mathcal{R}b \Rightarrow a \in B$. The upper closure of a subset of $(A, \mathcal{R})$ is the smallest upset of $(A, \mathcal{R})$ containing it: $\forall B \subseteq A, cl_A^{\mathcal{R}}(B) := \{a \in A : \exists b \in B \; a\mathcal{R}b\}$.*

We postulate that the process is bounded by two assumptions of rationality, individual and collective.

- At the individual level, for all points of view $i \in \mathcal{N}$, the approved subset of alternatives $\mathcal{A}_i \subseteq \mathbb{X}$ should be an upset for the preference relation $\succsim_i$. Hence, there is no pair of alternatives $x, x' \in \mathbb{X}$ where x is preferred to x' w.r.t. $\succsim_i$, x' is approved by i but not x.
- At the collective level, an alternative $x \in \mathbb{X}$ is collectively approved and sorted into the upper category if, and only if, it is approved by a sufficient coalition of points of view. We assume the set of sufficient coalitions $\mathcal{S} \subseteq \mathcal{P}(\mathcal{N})$ is fixed, and is an upset for inclusion. Hence, if a coalition is sufficient, any superset of this coalition is also sufficient (and if a coalition is insufficient, any subset of it is also insufficient). We do not assume the set of sufficient coalitions has an additive structure, as opposed to weighted voting games or approval balloting [Laslier and Sanver, 2010].

These two stages form the noncompensatory sorting model:

Definition 2 (NCS - noncompensatory sorting model, [Bouyssou and Marchant, 2007])**.** *Given a set of alternatives $\mathbb{X}$, a set of points of view $\mathcal{N}$, and a tuple of complete preorders $\succsim_i, i \in \mathcal{N}$, if $\mathcal{S}$ is an upset of $(\mathcal{P}(\mathcal{N}), \subseteq)$ and a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$,* [1] *the noncompensatory sorting model with parameters $(\mathcal{S}, \langle \mathcal{A}_i \rangle)$ is the function $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ mapping alternatives from $\mathbb{X}$ to categories in $\{\text{GOOD}, \text{BAD}\}$ such that the alternative x is assigned to the upper category* GOOD *if, and only if, the set of points of view according to which x is approved is sufficient,* i.e.

$$NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(x) = \begin{cases} \text{GOOD}, \text{ if } \{i \in \mathcal{N} : x \in \mathcal{A}_i\} \in \mathcal{S} \\ \text{BAD}, \text{ else} \end{cases}$$

$\mathcal{S}$ is the set of sufficient coalitions *of the model, and each $\mathcal{A}_i$ is the* approved set *according to the point of view $i \in \mathcal{N}$.*

Example 2. *(ex. 1 continued) Suppose the approved sets are as follows: $\mathcal{A}_1 := \{a, b, f\}, \mathcal{A}_2 := \{e, b, c\}, \mathcal{A}_3 := \{f, a, b\}, \mathcal{A}_4 := \{d, a, c\}, \mathcal{A}_5 := \{c, e, b\}$, corresponding to the three best alternatives according to the respective points of view (3-approval). Suppose also the points of view are aggregated according to the simple majority rule, i.e. $B \in \mathcal{S} \iff |B| \geq 3$. Then, the corresponding noncompensatory model assigns a, b, c to the* GOOD *category, and d, e, f to the* BAD *one. Hence, $\alpha := \{(a, \text{GOOD}), (b, \text{GOOD}), (c, \text{GOOD}), (d, \text{BAD}), (e, \text{BAD}), (f, \text{BAD})\}$. We note the same assignment α can be obtained with different sorting parameters, e.g. approved sets $\mathcal{A}'_1 := \{a, b, f\}, \mathcal{A}'_2 := \{e, b, c, d, a\}, \mathcal{A}'_3 := \{\}, \mathcal{A}'_4 := \{d, a, c\}, \mathcal{A}'_5 := \{c\}$ and sufficient coalitions $\mathcal{S}'$ containing the coalitions $\{1, 2\}, \{5\}$ and their supersets.*

This model may appear particularly unwieldy to use explicitly, as it requires to handle a set of sufficient coalitions that lies in the power set of the points of view.

[1] Meaning $\langle \mathcal{A}_i \rangle_{i \in \mathcal{N}}$ is a tuple of subsets of $\mathbb{X}$ such that, for all $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$. Also, throughout the paper, when the indexing is left unspecified, the tuples are indexed by points of view $i \in \mathcal{N}$.

We propose an indirect approach w.r.t. the parameters of the noncompensatory sorting model implicitly describing the decision process: we suppose the inputs (ordinal preferences over the alternatives according to each point of view) and outputs (an assignment of each alternative to a category, either GOOD or BAD) of the aggregation model are given, and we query the parameters (sufficient coalitions of points of view and accepted sets according to each point of view) of the model. Unlike the usual learning approach, based on the inverse problem of finding the *value* of a suitable tuple of parameters permitting to restore the output given the input, we instead focus on versions of this problem where the issue is merely the *existence* of such a tuple of parameters, and, in the case of a positive answer, to find suitable values for the accepted sets (but not for the set of sufficient coalitions).

Definition 3 (Inverse noncompensatory sorting problem: Inv-NCS)**.** *Given an assignment $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ of alternatives to categories, we say that α* can be represented in the noncompensatory sorting model *if, and only if, there is a pair of parameters $(\mathcal{S}, \langle \mathcal{A}_i \rangle)$ where $\mathcal{S}$ is an upset of $(\mathcal{P}(\mathcal{N}), \subseteq)$ and $\langle \mathcal{A}_i \rangle_{i \in \mathcal{N}}$ is a tuple of subsets of $\mathbb{X}$ such that, for all $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$, so that $\alpha \equiv NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$.*

We say that α is a *possible assignment* if it is a YES instance of Inv-NCS, *i.e.* α can be represented in the noncompensatory sorting model. When there is some jurisprudence $\alpha^\star$, the assignment of a new candidate x can be *necessary*, in the sense that no other assignment is possible.

Definition 4 (Necessary assignment w.r.t. reference cases)**.** *Given a* YES *instance $\alpha^\star$ of Inv-NCS, an alternative $x \in \mathbb{X}$ is necessarily assigned to a category $C \in \{\text{GOOD}, \text{BAD}\}$ w.r.t. assignment $\alpha^\star$ if $\alpha^\star \cup \{(x, \overline{C})\}$ is a* NO *instance of Inv-NCS, where $\overline{C}$ denotes the category opposite to C.*

3 Feasibility of the Inverse NCS Problem

In this section, we propose a characterization of the possibility, given ordinal preferences over the alternatives according to each point of view and an assignment of each alternative to a category, either GOOD or BAD, of representing this assignment in the non-compensatory sorting model. This formulation circumvents any reference to the power set of points of view, so we derive a compact SAT formulation for the inverse problem, which is shown to be NP-hard.

3.1 Inv-NCS with Fixed Approved Sets

When the approved sets are given, solving the inverse NCS problem – *i.e.* learning a set of sufficient coalitions permitting to represent the assignment in the noncompensatory sorting model – is similar to learning a disjunctive normal form from training examples. From this observation, we derive a tractable (computable in polynomial time) algorithm yielding the *version space* [Mitchell, 1982] of the noncompensatory sorting model with fixed approved sets:

Definition 5 (Observed sufficient and insufficient coalitions given approved sets)**.** *Given $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ and a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$, we note:*

$$\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^{\supseteq}_{\mathcal{P}(\mathcal{N})}\left(\bigcup_{g \in \alpha^{-1}(\text{GOOD})} \{i \in \mathcal{N} : g \in \mathcal{A}_i\}\right),$$

$$\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) := cl^{\subseteq}_{\mathcal{P}(\mathcal{N})}\left(\bigcup_{b \in \alpha^{-1}(\text{BAD})} \{i \in \mathcal{N} : b \in \mathcal{A}_i\}\right)$$

Proposition 1 (Lower and upper bounds for the sufficient coalitions given the approved sets)**.** *Given an assignment α, a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$ and an upset $\mathcal{S}$ of $(\mathcal{P}(\mathcal{N}), \subseteq)$, α is represented by the noncompensatory sorting model $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ if, and only if:*

$$\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \subseteq \mathcal{S} \subseteq \mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$$

Proof. α is represented by $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ iff i) for all alternatives $g \in \alpha^{-1}(\text{GOOD}), NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(g) = \text{GOOD}$; and ii) for all alternatives $b \in \alpha^{-1}(\text{BAD}), NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}(b) = \text{BAD}$

i) holds iff $\mathcal{S}$ contains $\bigcup_{g \in \alpha^{-1}(\text{GOOD})} \{i \in \mathcal{N} : g \in \mathcal{A}_i\}$ and, as a consequence of being an upset for inclusion, $\mathcal{S}$ contains $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$. ii) holds iff $\mathcal{S}$ does not contain any coalition pertaining neither to $\bigcup_{b \in \alpha^{-1}(\text{BAD})} \{i \in \mathcal{N} : b \in \mathcal{A}_i\}$ nor to $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$. □

Corollary 1 (complexity of Inv-NCS with fixed approved sets)**.** *Given an assignment α of alternatives to categories and a tuple $\langle \mathcal{A}_i \rangle$ of upsets of $\langle (\mathcal{P}(\mathbb{X}), \succsim_i) \rangle$, the problem of deciding whether α can be represented in the noncompensatory sorting model with approved sets $\langle \mathcal{A}_i \rangle$ is tractable (computable in polynomial time).*

Indeed, it boils down to checking whether $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ is empty or not, which is $O(|\mathbb{X}|^2 \cdot |\mathcal{N}|)$.

3.2 A Pairwise Formulation for Inv-NCS

The following Theorem is very important as it says that, in order to check that an assignment α is compatible with NCS, it is equivalent to find approval subsets over each point of view such that one can discriminate each pair of GOOD and BAD alternatives on at least one point of view (*i.e.* the GOOD alternative is approved on this point of view, and not the BAD one). Interestingly, the concept of sufficient coalitions disappears in the characterization.

Theorem 1 (Pairwise formulation of the noncompensatory sorting model)**.** *An assignment α of alternatives to categories can be represented in the noncompensatory sorting model if, and only if, there is a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{\mathcal{N}}$ such that:*

1. *for each point of view $i \in \mathcal{N}$, $\mathcal{A}_i$ is an upset of $(\mathbb{X}, \succsim_i)$*
2. *for each pair of alternatives $(g, b) \in \alpha^{-1}(\text{GOOD}) \times \alpha^{-1}(\text{BAD})$, there is at least one point of view $i \in \mathcal{N}$ such that $g \in \mathcal{A}_i$ and $b \notin \mathcal{A}_i$.*

Proof. [¬(1+2) ⇒ ¬NCS] If there are two alternatives $g \in \alpha^{-1}(\text{GOOD})$ and $b \in \alpha^{-1}(\text{BAD})$ that falsify Condition 2, then, for any potential parameters $\mathcal{S}, \langle \mathcal{A}_i \rangle$ of a noncompensatory sorting model, the nesting $\{i \in \mathcal{N} : g \in \mathcal{A}_i\} \subseteq \{i \in \mathcal{N} : b \in \mathcal{A}_i\}$ results in a sorting $NCS_{\mathcal{S}, \langle \mathcal{A}_i \rangle}$ at least as favorable to b as to g, whereas $\alpha(b) = \text{BAD}$ is strictly worse than $\alpha(g) = \text{GOOD}$.

[(1+2) ⇒ NCS] Given a tuple $\langle \mathcal{A}_i \rangle \in \mathcal{P}(\mathbb{X})^{\mathcal{N}}$ satisfying conditions 1 and 2, we consider the sets of coalitions $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$.

According to Proposition 1, α can be represented in the noncompensatory model iff $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) = \emptyset$. Suppose this intersection is nonempty, and let $B \in \mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$. By definition of $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$, there is an alternative $g \in \alpha^{-1}(\text{GOOD})$ such that $B \supseteq \{i \in \mathcal{N} : g \in \mathcal{A}_i\}$: for all points of view $i \notin B$, $g \notin \mathcal{A}_i$. By definition of $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$, there is an alternative $b \in \alpha^{-1}(\text{BAD})$ such that $B \subseteq \{i \in \mathcal{N} : b \in \mathcal{A}_i\}$: for all points of view $i \in B$, $b \in \mathcal{A}_i$. Consequently, there is no point of view according to which g is accepted but not b, contradicting condition 2. Hence, $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha) \cap \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha) = \emptyset$. □

3.3 Complexity of Inv-NCS

We show that the inverse NCS problem is intractable.

Proposition 2 (NP-hardness of Inv-NCS)**.**

Given an assignment α of alternatives to categories, the problem of deciding whether α can be represented in the noncompensatory sorting model is NP-hard.

Proof. By reduction from SAT: consider a SAT instance in conjunctive normal form, with n variables $y^1, \ldots, y^n$ and m clauses $c_1 \wedge \cdots \wedge c_m$. We build a gadget assignment with $m+n$ points of view and $2m$ alternatives: $g_1, \ldots, g_m$ are assigned to GOOD whereas $b_1, \ldots, b_m$ are assigned to BAD. First, let us focus on the first m points of view: for each $k \in 1 \ldots m$, let $g_k \sim_k b_k \succ_k g_1 \sim_k \cdots \sim_k g_{k-1} \sim_k g_{k+1} \sim_k \cdots \sim_k g_m \sim_k b_1 \sim_k \cdots \sim_k b_{k-1} \sim_k b_{k+1} \sim_k \cdots \sim_k b_m$. The preference $\succsim_k$ has two equivalence classes, the upper one containing $\{g_k, b_k\}$ and the lower one containing $\bigcup_{k' \neq k} \{g_{k'}, b_{k'}\}$. The n last points of view of the gadget are built considering the SAT formula. From the j-th clause, written in disjunctive form $c_j := \bigvee_{k \in P_j} y^k \vee \bigvee_{k \in N_j} \neg y^k$, where P_j and N_j are disjoint subsets of $1 \ldots n$ indexing the positive (resp. negative) atoms of c_j, we build the preference relation $\succsim_{j+m}$. It has at most 3 equivalence classes: the uppermost containing the alternatives $\bigcup_{k \in P_j} \{g_k\}$, the one in the middle containing $\bigcup_{k \in P_j} \{b_k\} \cup \bigcup_{k \in N_j} \{g_k\}$, and the lowest containing $\bigcup_{k \in N_j} \{b_k\} \cup \bigcup_{k \notin P_j \cup N_j} \{g_k, b_k\}$. We note trivial accepted sets – *i.e.* points of view $i \in \mathcal{N}$ such that $\mathcal{A}_i = \emptyset$ or $\mathcal{A}_i = \mathbb{X}$ – do not contribute to the feasibility of the inverse NCS problem. For the m first points of view, there is only one nontrivial accepted set: it accepts the upper class and rejects the lower one. For the n last points of view of the gadget, the nontrivial accepted sets accept the uppermost equivalence class, reject the lowest class, and either accept or reject the class in the middle. We define a one-to-one mapping between the nontrivial accepted sets of the gadget and the assignment of the n variables of the SAT problem: y^j is False $\iff \bigcup_{k \in P_j} \{b_k\} \cup \bigcup_{k \in N_j} \{g_k\} \in \mathcal{A}_{m+j}$. Each non trivial assignment discriminates all pairs $(g_k, b_{k'})$ with $k \neq k'$ w.r.t. the point of view k. The pairs (g_k, b_k) is discriminated iff the clause c_k is satisfied. Thus, a solution of the SAT problem is mapped to a tuple of accepted sets that discriminates all pairs with opposite assignments and reciprocally. □

3.4 A Compact SAT Formulation for Inv-NCS

We leverage Theorem 1 by formulating a boolean satisfiability problem that answers the decision problem: can the assignment α be represented in the non-compensatory model? If the instance is a YES, any solution of the satisfiability problem translates into suitable, yet arbitrary, explicit values for the approved sets. Upper and lower bounds for the set of sufficient coalitions can be obtained thanks to Proposition 1.

Corollary 2 (CNF Pairwise SAT formulation for NCS)**.** *Let $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ an assignment. We define the boolean function $\phi_\alpha^{pairwise}$ with variables:*

- *$\lambda_{i,x}$ indexed by a point of view $i \in \mathcal{N}$, and a value $x \in \mathbb{X}$,*
- *$\mu_{i,g,b}$ indexed by a point of view $i \in \mathcal{N}$, a* good *alternative $g \in \alpha^{-1}(\text{GOOD})$ and a* bad *alternative $b \in \alpha^{-1}(\text{BAD})$,*

as the conjunction of clauses: $\phi_\alpha^{pairwise} := \phi_\alpha^1 \wedge \phi_\alpha^2 \wedge \phi_\alpha^3 \wedge \phi_\alpha^4$

$$\phi_\alpha^1 := \bigwedge_{i \in \mathcal{N}} \bigwedge_{x' \succsim_i x} (\lambda_{i,x'} \vee \neg \lambda_{i,x})$$

$$\phi_\alpha^2 := \bigwedge_{i \in \mathcal{N},\, g \in \alpha^{-1}(\text{GOOD}),\, b \in \alpha^{-1}(\text{BAD})} (\neg \mu_{i,g,b} \vee \neg \lambda_{i,b})$$

$$\phi_\alpha^3 := \bigwedge_{i \in \mathcal{N},\, g \in \alpha^{-1}(\text{GOOD}),\, b \in \alpha^{-1}(\text{BAD})} (\neg \mu_{i,g,b} \vee \lambda_{i,g})$$

$$\phi_\alpha^4 := \bigwedge_{g \in \alpha^{-1}(\text{GOOD}),\, b \in \alpha^{-1}(\text{BAD})} (\bigvee_{i \in \mathcal{N}} \mu_{i,g,b})$$

α can be represented in the noncompensatory sorting model if, and only if, $\phi_\alpha^{pairwise}$ is satisfiable.

Moreover, if $\langle \lambda_{i,x} \rangle, \langle \mu_{i,g,b} \rangle$ is an antecedent of 1 by $\phi_\alpha^{pairwise}$, then the noncompensatory sorting model $NCS_{\mathcal{S},\langle \mathcal{A}_i \rangle}$ with accepted sets defined by $\mathcal{A}_i := \{x \in \mathbb{X} : \lambda_{i,x} = 1\}$ and any upset $\mathcal{S}$ of $(\mathcal{P}(\mathcal{N}), \subseteq)$ of sufficient coalitions containing the upset $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$ and disjoint from the lower set $\mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ satisfies $\alpha \equiv NCS_{\mathcal{S},\langle \mathcal{A}_i \rangle}$.

Variables $\lambda_{i,x}$ are assigned to 1 when the alternative x is accepted from the point of view i, and variables $\mu_{i,g,b}$ are assigned to 1 when the point of view i accepts g but not b.

The clauses ϕ_α^1 ensure the sets of accepted values of each point of view meet the first condition of Theorem 1, i.e. $\mathcal{A}_i$ is an upset. The clauses ϕ_α^2 (resp. ϕ_α^3) ensure each variable $\mu_{i,g,b}$ cannot take a value of one unless g is accepted (resp. unless b is not accepted). The clauses ϕ_α^4 ensure the second condition of Theorem 1 is met.

The formulation is compact: $O(|\mathcal{N}| \cdot |\mathbb{X}|^2)$ variables, $O(|\mathcal{N}| \cdot |\mathbb{X}|^2)$ binary clauses and $O(|\mathbb{X}|^2)$ $|\mathcal{N}|$-ary clauses.

4 Accountable Decisions with Inv-NCS

In this section we describe how the theoretical and algorithmic tools described in Section 3 in order to assess the feasibility of the inverse NCS problem (see Def. 3) can be used to support a decision process. More precisely, we address the situation described in Section 1 where a committee has to assign alternatives either to the GOOD or the BAD category, and to account for this assignment. Section 4.1 addresses the first situation S1, where an audit is commissioned to check the compliance of the committee to its terms of reference, by referring to the notion of *possible* assignment. Section 4.2 addresses the second situation S2, where the committee is challenged by a stakeholder to defend a specific decision, by referring to the notion of *necessary* assignment.

4.1 Auditing Conformity

We consider the situation S1 depicted in Section 1, where an independent audit agency has to check that the decision α of the committee on candidates $\mathbb{X}$ is compatible with NCS. We assume $\mathbb{X}^\star = \emptyset$: all the assignments should be justified together, and none should be taken for granted.

Should the burden of proof be left to the auditor, the audit procedure could require either i) full disclosure of the preference profile $\langle(\mathbb{X}, \succsim_i)\rangle_{i\in\mathcal{N}}$, and the auditor solving the NP-hard Inv-NCS problem, e.g. using a SAT solver and Corollary 2; or ii) full disclosure of the approved sets $\langle\mathcal{A}_i\rangle_{i\in\mathcal{N}}$, and the auditor solving the tractable Inv-NCS with fixed accepted sets problem as described by Proposition 1.

If we consider putting the burden of proof on the committee, Theorem 1 can be leveraged to compute and provide a certificate of feasibility for Inv-NCS(α) that involves the disclosure of less information, as illustrated below:

Example 3. *(ex. 2 cont.) If the approved sets of the committee are $\mathcal{A}_1, \ldots, \mathcal{A}_5$, then it needs to disclose information concerning three points of view in order to prove the assignment α is consistent with an approval procedure, e.g. :*

- *according to the first point of view, b is approved (and so is a which is better than b) whereas e is not (and neither is d which is worse than e), hence the procedure is able to discriminate a, b from d, e;*
- *according to the second point of view, c is approved (and so is b which is better than c) whereas d is not (and neither is f which is worse than d), hence the procedure is able to discriminate b, c from d, f;*
- *according to the fourth point of view, c is approved (and so is a which is better than c) whereas e is not (and neither is f which is worse than e), hence the procedure is able to discriminate a, c from e, f.*

The following table summarizes the points of view permitting to discriminate each pair:

		BAD		
		d	*e*	*f*
GOOD	*a*	*1*	*1*	*4*
	b	*1*	*1*	*2*
	c	*2*	*4*	*2*

This manner of arguing that a given assignment is indeed a possible outcome of an approval sorting procedure can be formalized into an *argument scheme*, an operator tying a tuple of premises – pieces of information satisfying some conditions – to a conclusion [Walton, 1996].

Definition 6 (Argument Scheme (AS1))**.** *We say a tuple $\langle(i_1, g_1, G_1, b_1, B_1), \ldots, (i_n, g_n, G_n, b_n, B_n)\rangle$ instantiates the argument scheme AS1 supporting the assignment α if: i) for all $k \in \{1 \ldots n\}$, $i_k \in \mathcal{N}$, $g_k \in G_k$, $\alpha(G_k) = \{\text{GOOD}\}$, $\forall g \in G_k, g \succsim_{i_k} g_k$, $b_k \in B_k$, $\alpha(B_k) = \{\text{BAD}\}$, $\forall b \in B_k, b_k \succsim_{i_k} b$ and $g_k \succ_{i_k} b_k$; and ii) $\bigcup_{k\in\{1\ldots n\}} G_k \times B_k = \alpha^{-1}(\text{GOOD}) \times \alpha^{-1}(\text{BAD})$*

Hence, according to the point of view i_k, g_k is the least preferred alternative in the subset of GOOD alternatives G_k and it is preferred to b_k, the most preferred alternative in the subset of BAD alternatives B_k. This scheme is somewhat frugal in the number of pairs of the profile $\langle(\mathbb{X}, \succsim_i)\rangle_{i\in\mathcal{N}}$ revealed to the auditor, as the comparisons inside $G_k \times G_k$ or $B_k \times B_k$ are not disclosed. Theorem 1 can be reworded as follows:

Corollary 3. *An assignment α is a* YES *instance of Inv-NCS if, and only if, there is an instance of AS1 supporting it.*

Example 4. *(Example 3 cont.) The explanations given in Example 3 instantiate AS1 as follows: $\big\langle(1, b, \{a, b\}, e, \{d, e\}), (2, c, \{b, c\}, d, \{d, f\}), (4, c, \{a, c\}, e, \{e, f\})\big\rangle$*

The length n of an explanation instantiating the argument scheme AS1 offers an indication regarding its cognitive complexity as well as the amount of information disclosed to the auditor. Therefore, we would rather provide the shortest possible explanations, and strive to mention as few points of view as possible. Obviously, an explanation needs to reference a specific point of view at most once, so $n \leq |\mathcal{N}|$. Unfortunately, the following result shows that one might require all points of view in a complete explanation, even in situations with relatively few alternatives.

Proposition 3. *For every set of points of view $\mathcal{N}$, there exists a set of $|\mathcal{N}| + 1$ alternatives $\mathbb{X}$ and an assignment $\alpha : \mathbb{X} \to \{\text{GOOD}, \text{BAD}\}$ for which any tuple instantiating the argument scheme AS1 and supporting α has length $|\mathcal{N}|$.*

Sketch of the Proof. The result is shown by induction on $|\mathcal{N}|$. For $|\mathcal{N}| = \{1\}$, we consider $\alpha_1 := \{(g, \text{GOOD}), (b, \text{BAD})\}$ with $g \succ_1 b$. Consider by induction an assignment α_p on p candidates $\mathbb{X}_p$ assessed on points of view $\mathcal{N} = \{1 \ldots p\}$. We introduce a new alternative z, judged as GOOD, and a new point of view $p + 1$, such that the candidates in $\mathbb{X}_p$ are indifferent on the new point of view, and z can be discriminated from b only on the new point of view. □

4.2 Justifying Individual Decisions

We now wish to justify the decision of the committee on a candidate $x \in \mathbb{X}$ (Situation S2). As we have seen in the previous section, a complete explanation of the assignment of x necessarily implies the disclosure of many information related to the other candidates, which might not be acceptable.

A possible solution is for committee to base their decision on reference cases, an assignment $\alpha^* : \mathbb{X}^* \to \{\text{GOOD}, \text{BAD}\}$, e.g. compiling past decisions that are representative of its functioning mode. In order to get rid of the influence of the other candidates, we are looking for *necessary assignments* given these reference cases.

Example 5. *(ex. 2 cont.) We consider the alternatives a, b, c, d, e, f and their assignment $\alpha^\star$ have a reference status, and we are interested in deciding on the assignment of two candidates, x, y such that:*

$$a \succ_1 f \succ_1 b \succ_1 e \succ_1 c \succ_1 y \succ_1 d \succ_1 x$$
$$e \succ_2 b \succ_2 y \succ_2 c \succ_2 d \succ_2 a \succ_2 f \succ_2 x$$
$$f \succ_3 a \succ_3 d \succ_3 b \succ_3 y \succ_3 x \succ_3 e \succ_3 c$$
$$d \succ_4 a \succ_4 c \succ_4 e \succ_4 x \succ_4 y \succ_4 f \succ_4 b$$
$$c \succ_5 y \succ_5 e \succ_5 b \succ_5 f \succ_5 x \succ_5 d \succ_5 a$$

It is not possible to represent the assignment (x, GOOD) together with the reference assignment α. Thus, x is necessarily assigned to BAD *. On the contrary, both assignments (y, GOOD) and (y, BAD) can be represented together with α.*

Necessary Decisions Entailed by the Jurisprudence

An explanation of the *necessity* of an assignment is intrinsically more complex than that for its *possibility*: one needs to prove that it is not possible to separate all pairs of GOOD and BAD candidates on at least one point of view. The proof relies on some deadlock that needs to be shown. Formally, this situation manifests itself in the form of an unsatisfiable boolean formula, e.g. given by Corollary 2. The unsatisfiability of the entire formula can be reduced to a $\subseteq$-minimal unsatisfiable subset of clauses (MUS), which are commonly used as certificates of infeasibility, and can also be leveraged to produce *explanations* [Junker, 2004; Besnard *et al.*, 2010; Geist and Peters, 2017]. In the case of the necessary decisions by approval sorting with a reference assignment, any MUS pinpoints a set of pairs of alternatives in $(\alpha^{-1}(\text{GOOD}) \cup \{x\}) \times \alpha^{-1}(\text{BAD})$ that cannot be discriminated simultaneously according to the points of view.

Example 6. *(ex. 5 cont.) Consider the subset of alternatives c, d, e, f, x, and assume x to be assigned to* GOOD*. Each pair in $GB := \{(c, e), (x, d), (x, f)\}$ needs to be discriminated from at least one point of view in $\mathcal{N}$, but this is not possible simultaneously: i) none of the pairs in GB can be discriminated neither from the first, the second nor the third point of view, as the overall* GOOD *alternative is deemed worse than the* BAD *one. ii) no more than one pair in GB can be discriminated according to each point of view among $\{4, 5\}$, and there are more pairs to discriminate than points of view.*

The pattern of deadlock illustrated by Example 6 can be generalized and formalized into an *argument scheme*, with *premises*: i) a k-tuple of pairs $\langle (g^1, b^1), \ldots, (g^k, b^k) \rangle$ of alternatives with opposite assignment, ii) a subset of points of view $B \subseteq \mathcal{N}$ with cardinality $k - 1$, such that, according to all points of view $i \notin B$, $b^j \succ_i g^j$ for all j, and, according to all points of view $i \in B$ the intervals $]b^1, g^1]_i, \ldots,]b^k, g^k]_i$ are pairwise disjoint.

Clearly, the existence of an argument instantiating the premises of this scheme is a sufficient condition for the infeasibility of representing the given assignment in the noncompensatory model, which in turn yields the *conclusion* that the candidate x is necessarily assigned to the other category.

If we assume that the cognitive burden demanded by an explanation along the lines of this argument scheme increases with the number of its premises, we derive an implicit hierarchy among the necessary decisions supported by the scheme, with a nesting $\mathcal{E}_1 \subseteq \mathcal{E}_2 \subseteq \cdots \subseteq \mathcal{E}_{|\mathcal{N}|+1}$, where $\mathcal{E}_k$ denotes the set of decisions supported by a scheme with premises referencing at most k pairs of alternatives with opposite assignment. $\mathcal{E}_1$ is exactly the set of decisions stemming from Pareto dominance, where a candidate is either at least as good as a reference alternative in the GOOD category, or at most as good as a reference alternative in the BAD category.

The question of deciding if this scheme captures a necessary condition, *i.e.* if any decision entailed by the jurisprudence can be supported by such an explanation, is left open.

Ambivalent Situations

It may happen that, for a given candidate, both assignments to GOOD and to BAD are possible. This situation is obviously all the more frequent as the reference set is small, or the number of points of view is high. In such a case, a design option would consist in constraining the decision of the committee, either favorably (e.g. following an *innocent unless proven guilty* principle) or unfavorably (e.g. following a *precautionary principle*). Another, more common, venue would give the freedom of choice to the committee. In this case, as opposed to the situation where the decision is entailed by the jurisprudence, and where the committee just needs to make obvious the link between the current case and the reference cases, the committee needs to disclose some information concerning its inner functioning. In some cases, though, Proposition 1 offers a solution that avoids a complete disclosure: suppose that, given the approved sets $\langle \mathcal{A}_i \rangle$, the candidate is approved from a coalition of points of view that is known to be insufficient (resp. sufficient), because a reference alternative is assigned to the BAD (resp. GOOD) category in a similar, or even better (resp. worse) situation than the candidate. This fortunate situation circumvents the need of discussing the particulars of the set of sufficient coalitions by referring to its upper bound $\mathcal{P}(\mathcal{N}) \setminus \mathcal{F}_{\langle \mathcal{A}_i \rangle}(\alpha)$ (resp. lower bound $\mathcal{T}_{\langle \mathcal{A}_i \rangle}(\alpha)$).

Example 7. *(ex. 6 cont.) According to the first point of view, y is disapproved, as it is worse than $c \notin \mathcal{A}_1$. According to the third point of view, y is disapproved, as it is worse than $b \notin \mathcal{A}_3$. According to the fifth point of view, y is disapproved, as it is worse than $f \notin \mathcal{A}_5$. Furthermore, being approved according to both the second and fourth points of view is not enough to warrant access to the* GOOD *category, as illustrated by e. Hence, y is assigned to the* BAD *category.*

5 Related Work and Conclusion

In this paper we are interested in the problem of accountability of decisions issued from a noncompensatory sorting model (NCS) [Bouyssou and Marchant, 2007]. Two situations have been mainly studied. In the first one, the committee needs to justify that its decision is a possible NCS assignment. A characterization result helps to turn the existence of such assignment to finding separations of the pairs of GOOD and BAD candidates over at least one point of view, which can be formulated as a SAT problem. This allows us to generate a single argument scheme that can explain all possible NCS assignments. The second situation arises when the assignment of a new candidate is necessarily derived from jurisprudence. Thanks to the characterization result, one can also construct an argument scheme representing deadlock situations. The use of argument schemes as formal tools to convey explanation in the context of multi-criteria aiding has also been advocated in [Labreuche, 2011; Nunes *et al.*, 2014; Belahcene *et al.*, 2017].

Our solutions stem from an original take of the dual notions of *possibility* and *necessity*, often used in so-called robust optimization, decision making [Greco *et al.*, 2010] or voting contexts [Boutilier and Rosenschein, 2016] to account for incomplete information, conveying epistemic stances of skepticism or credulousness. Instead we use them to describe the leeway left to the committee in setting its expectations: the decisions taken are bound from above by possibility, described as the feasibility of the Inv-NCS problem related to

their decision, and from below by necessity, described as the infeasibility of the Inv-NCS problem simultaneously related to the reference cases and impossible assignments.

Barrot *et al.* (2013) study the problem of identifying the *possible* winners of an approval election, when votes are given but approval thresholds are unspecified. They show that determining whether a set of candidates are co-winners is NP-complete when voters have fixed (even equal) importance. Approval voting has been studied in the context of multi-winner elections [Aziz *et al.*, 2015], which may seem close to our setting: indeed, we could see the candidates ranked in GOOD as the winners. However, in our context, each candidate is ranked without consideration to the other candidates, and voters are not assumed to have equal importance.

Finally, several algorithms have been proposed to learn the parameters of a noncompensatory sorting model from observation: [Leroy *et al.*, 2011] relies on a MIP formulation, [Sobrie *et al.*, 2015] relies on a metaheuristic.

Acknowledgments

This work is partially supported by the ANR project 14-CE24-0007-01 - CoCoRICo-CoDec.

References

[Aziz *et al.*, 2015] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Simon Mackenzie, Nicholas Mattei, and Toby Walsh. Computational aspects of multi-winner approval voting. In *Proceedings of AAMAS*, pages 107–115, 2015.

[Barrot *et al.*, 2013] Nathanaël Barrot, Laurent Gourvès, Jérôme Lang, Jérôme Monnot, and Bernard Ries. Possible winners in approval voting. In *Proceedings of the third International conference on Algorithmic Decision Theory*, pages 57–70, 2013.

[Belahcene *et al.*, 2017] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. A model for accountable ordinal sorting. In *Proceedings of the 26 International Joint Conference on Artificial Intelligence*, pages 814–820, 2017.

[Besnard *et al.*, 2010] Philippe Besnard, Éric Grégoire, Cédric Piette, and Badran Raddaoui. MUS-based generation of arguments and counter-arguments. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 239–244, 2010.

[Boutilier and Rosenschein, 2016] Craig Boutilier and Jeffrey S. Rosenschein. *Incomplete Information and Communication in Voting*, page 223–258. Cambridge University Press, 2016.

[Bouyssou and Marchant, 2007] Denis Bouyssou and Thierry Marchant. An axiomatic approach to noncompensatory sorting methods in MCDM, i: The case of two categories. *EJOR*, 178(1):217–245, 2007.

[Bouyssou *et al.*, 2006] Denis Bouyssou, Thierry Marchant, Marc Pirlot, Alexis Tsoukiàs, and Philippe Vincke. *Evaluation and decision models with multiple criteria: Stepping stones for the analyst*. International Series in Operations Research and Management Science. Springer, 2006.

[Doshi-Velez *et al.*, 2017] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of AI under the law: The role of explanation. *CoRR*, abs/1711.01134, 2017.

[Geist and Peters, 2017] Christian Geist and Dominik Peters. Computer-aided methods for social choice theory. In Ulle Endriss, editor, *Trends in Computational Social Choice*, chapter 13, pages 249–267. AI Access, 2017.

[Greco *et al.*, 2010] Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, 207(3):1455 – 1470, 2010.

[Junker, 2004] Ulrich Junker. Quickxplain: Preferred explanations and relaxations for over-constrained problems. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pages 167–172, 2004.

[Kroll *et al.*, 2017] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017.

[Labreuche, 2011] Christophe Labreuche. A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence Journal*, 175:1410–1448, 2011.

[Laslier and Sanver, 2010] Jean-François Laslier and M. Remzi Sanver. *Handbook on Approval Voting*. Studies in Choice and Welfare. Springer, Boston, 2010.

[Leroy *et al.*, 2011] Agnes Leroy, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a multiple criteria sorting method. In *Proceedings of the second International conference on Algorithmic Decision Theory*, pages 219–233, 2011.

[Mitchell, 1982] Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[Nunes *et al.*, 2014] Ingrid Nunes, Simon Miles, Michael Luck, Simone Diniz Junqueira Barbosa, and Carlos José Pereira de Lucena. Pattern-based explanation for automated decisions. In *Proceedings of 21st ECAI*, pages 669–674, 2014.

[Sobrie *et al.*, 2015] Olivier Sobrie, Vincent Mousseau, and Marc Pirlot. Learning the parameters of a non compensatory sorting model. In *Proceedings of the fourth International Conference on Algorithmic Decision Theory*, volume 9346, pages 153–170, 2015.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99, May 2017.

[Walton, 1996] Douglas Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.