On Middle Grounds for Preference Statements

Anne-Marie George , Ana Ozaki
University of Oslo, Norway
{annemage, anaoz}@uio.no

Abstract

In group decisions or deliberations, stakeholders are often confronted with conflicting opinions. We investigate a logic-based way of expressing such opinions and a formal general notion of a middle ground between stakeholders. Inspired by the literature on preferences with hierarchical and lexicographic models, we instantiate our general framework to the case where stakeholders express their opinions using preference statements of the form *I prefer 'a' to 'b'*, where *'a'* and *'b'* are alternatives expressed over some attributes, e.g., in a trolley problem, one can express *I prefer to save 1 adult and 1 child to 2 adults (and 0 children)*. We prove theoretical results on the existence and uniqueness of middle grounds. In particular, we show that, for preference statements, middle grounds may not exist and may not be unique. We also provide algorithms for deciding the existence and finding middle grounds.

1 Introduction

High stake decisions or moral dilemmas, such as medical triage or the trolley problem, may prompt stakeholders to have strong opinions with little flexibility. The need to solve such decisions in real life requires the deliberation and consolidation of such possibly conflicting opinions. In this paper, we aim to break down stakeholder statements (e.g. statements about their moral preferences) into an agreeable set of statements — a *middle ground*. Efforts in defining such a notion of a middle ground have recently been made by Ozaki *et al.* (2024). However, their notion is designed for Horn logic. We propose a notion of middle ground for a generic logic formalized as a satisfaction system [Aiguier *et al.*, 2018] and provide a case study for a logic that expresses *preferences*.

Finding a middle ground between stakeholders can be an important first step to understanding and creating solutions for conflicting opinions. Applications of our work are thus manifold. Freedman *et al.* (2020), e.g., investigate human values in kidney exchanges, where patients are described by features of age, health, and drinking behaviour. Unsurprisingly, the 289 participants of their survey did not agree on the prioritisation of patients. Many other real-life scenarios may provoke conflicting opinions or values: end-of-life medical decisions, decisions prompting trade-offs between economic advantages and preservation of nature, or hiring where the roles in a hiring committee warrant emphasis of different applicant features.

The application scenarios above often include stakeholders who express preferences over alternatives. We concentrate our case study on satisfaction systems similar to those described by Wilson *et al.* (2015), with a language of comparative statements of the form "*I prefer a to b.*", where alternatives a and b are vectors of values from given variable domains. Models are then lexicographic or hierarchical orders, i.e., total pre-orders on the set of alternatives. That is, we assume stakeholders have (unknown) orders of importance for the features, by which they compare alternatives. These satisfaction systems transfer well to the Moral Machine Experiment [Awad *et al.*, 2018].

Example 1. *In the Moral Machine Experiment, participants (stakeholders) are asked to choose one out of two groups of individuals (alternatives) to save from a car accident. The participant's choices can be interpreted as comparative preference statements like "I prefer saving 1 adult, 4 children, and 0 dogs, to saving 2 adults, 3 children, and 3 dogs.", in symbols,* $(1, 4, 0) > (2, 3, 3)$. *This could, e.g., be modelled by the lexicographic model* $(\mathsf{child}, \mathsf{adult}, \mathsf{dog})$*, which prioritizes children, over adults, and adults over dogs, or by the hierarchical model* $(\{\mathsf{adult}, \mathsf{child}\}, \mathsf{child})$ *where alternatives are first compared on the number of humans, and only if they are equal (there are 5 humans in both groups), is the number of children considered (4 in the first group, 3 in the second). The number of dogs is disregarded in the second model.*

Inspired by these scenarios, this paper contributes with the following theoretical results.

General Notion of Middle Ground (MG): We provide a general definition of *middle ground* for satisfaction systems (Section 3.1), show conditions for existence of a MG (Section 3.2) and an algorithm for construction (Section 3.3).

Case Study for Preference Statements: We describe a satisfaction system similar to that of Wilson *et al.*(2015) for modelling preferences (Section 4.1), prove that existence and uniqueness of a MG is not guaranteed under this system (Section 4.2), and complexity results of deciding the consistency of preferences and existence of a MG (Section 4.3) for hierarchical models and the special case of lexicographic models.

Section 5 concludes. More proof details and discussions can be found in a longer version on Arxiv under the same title.

2 Related Work

There has been several logic-based approaches exploring the task of aggregating information and resolving conflicts in different fields such as non-monotonic reasoning [Horty, 1994; Delgrande and Schaub, 1997], belief merging [Gärdenfors, 1986], argumentation [Liao *et al.*, 2023], ontology repair [Moodley *et al.*, 2011], and normative reasoning in ethical and legal contexts [Ju *et al.*, 2020; Kollingbaum *et al.*, 2008].

Our work is most similar to that by Ozaki *et al.* (2024) which defines a middle ground notion for Horn logic and considers the Moral Machine Experiment [Awad *et al.*, 2018]. However, while the postulates (P'1-P'6) in their definition explicitly use structural aspects of Horn expressions like the antecedent and consequent, we facilitate a more general definition with postulates (P1-5) for satisfaction systems. When interpreting their notion of coherence as the counterpart of consistency, then the two liken each other in spirit. The middle ground is a set of statement that is in it self coherent / consistent (P'1 / P1), if possible equivalent to the union of all stakeholders' statements (P'2 / P2) and otherwise at least not in direct opposition to a stakeholders statement (P'3 / P3). Further, all statements in the middle ground should be motivated by stakeholder statements and retain their statements as close as possible (P'4-6 / P4-5). The work by Konieczny and Pérez (2011) in belief merging contains some postulates that resemble the notion of a middle ground. There, an operator takes possibly conflicting beliefs from multiple sources as input and returns the belief base that is closest to the input and some integrity constraints. This differs from middle ground in that they require more properties to hold. In particular, integrity constraints expressed in propositional logic need to be satisfied and the operator needs to compute *the* closest belief base (which may not exist or be unique for middle ground). Within social choice theory, Botan *et al.* (2023) investigate egalitarianism in judgement aggregation using propositional logic. Adler (2016) considers preference aggregation, arguing that preferences are more suitable than judgment for moral aggregation. As a fundamental difference, a middle ground might be insufficient on its own for subsequent decision making but maintains some agreement of all stakeholders that a compromise or aggregation found by means of social choice methods cannot facilitate.

Previous attempts to model preferences include weighted sums over features (which are restrictive w.r.t. to the nature of such features)[Wilson and Montazery, 2016], Pareto models which lead to only partial orders [George and Wilson, 2016], and perhaps most convincingly but also less tractable Conditional Preference Networks [Boutilier *et al.*, 2004] and the expressive prototypical preference logic [Bienvenu *et al.*, 2010]. Here, we lean our case study of preference statements onto the satisfaction systems described in [Wilson *et al.*, 2015]. Preferences are modelled by some kind of hierarchical models which are represented by importance orders on variables/ features of alternatives. One drawback of these models is that they require variables to be non-repeating in the importance order. Instead, we consider models that are non-empty and allow for repeating variables at several importance levels.

3 Middle Grounds for Sets of Statements

In this section we consider a general notion of middle ground for sets of statements and establish sufficient conditions for its existence. To make the presentation as general as possible, we first recall the notion of a *satisfaction system* [Aiguier *et al.*, 2018; Delgrande *et al.*, 2018; Guimarães *et al.*, 2023].

Definition 1 (Satisfaction System). *A satisfaction system is a triple* $(\mathcal{L}, \models, \mathfrak{M})$*, where* $\mathcal{L}$ *is a language,* $\mathfrak{M}$ *a set of models, and* $\models$ *a satisfaction relation on* $\mathfrak{M} \times \mathcal{L}$*. The relation* $\models$ *contains pairs of the form* (π, ϕ) *with model* π satisfying ϕ*.*

Given $\Phi \subseteq \mathcal{L}$*,* $\pi \models \Phi$ *iff* $\pi \models \phi$ *for all* $\phi \in \Phi$*. Given* $\Phi, \Phi' \subseteq \mathcal{L}$*, we say that* Φ entails Φ'*, written* $\Phi \models \Phi'$ *if, for all* $\pi \in \mathfrak{M}$*,* $\pi \models \Phi$ *implies* $\pi \models \Phi'$*. Let* $\mathsf{mod}(\Phi)$ *denote the set of models that satisfy* $\Phi \subseteq \mathcal{L}$*.*

Satisfaction systems have the following properties [Aiguier *et al.*, 2018]: if $\Phi \subseteq \Phi'$ then (1) $\mathsf{mod}(\Phi') \subseteq \mathsf{mod}(\Phi)$; and (2) $\Phi' \models \Phi$ (monotonicity). In this work, we consider satisfaction systems with *finite* $\mathcal{L}$. We may treat ϕ in $\mathcal{L}$ and singleton set $\{\phi\}$ interchangeably. Elements of $\mathcal{L}$ are called *statements*. A set of statements $\Phi \subseteq \mathcal{L}$ is *consistent* if $\mathsf{mod}(\Phi) \neq \emptyset$ and *falsifiable* if $\mathsf{mod}(\Phi) \neq \mathfrak{M}$. Also, Φ is *non-trivial* if it is consistent and falsifiable.

Throughout this section, we consider an arbitrary satisfaction system $(\mathcal{L}, \models, \mathfrak{M})$ and omit explicit references to it.

3.1 Notion of Middle Ground

Before we provide a formal definition of middle ground, we motivate it by considering a scenario where stakeholders have conflicting statements. Recall Example 1, suppose another participant prefers the second alternative, that is, to save 2 adults, 3 children, and 3 dogs, in symbols, $(1, 4, 0) < (2, 3, 3)$. Is there a middle ground for these two participants? The union of their preferences is clearly inconsistent but perhaps by "weakening" the second alternative, e.g., to $(2, 3, 0)$ and also making the preference of the first participant non-strict, we can find an agreeable statement. That is, intuitively, $(1, 4, 0) \geq (2, 3, 0)$ is "between" the preferences of both participants. This intuition is what we aim at capturing with middle grounds.

Definition 2 (Middle Ground). *Let* $\Phi_1, \ldots, \Phi_n$ *be non-trivial sets of statements, each associated with a stakeholder* $i \in \{1, \ldots, n\}$*. A set of statements* Φ *is a* **middle ground** *for* $\Phi_1, \ldots, \Phi_n$ *if it satisfies each of the following postulates:*

(P1) Φ *is non-trivial;*

(P2) if $\bigcup_{i=1}^{n} \Phi_i$ *is consistent, then* $\Phi \equiv \bigcup_{i=1}^{n} \Phi_i$*;*

(P3) for each $\phi \in \Phi$ *and for all* $i \in \{1, \ldots, n\}$ *and all* $\phi_i \in \Phi_i$*, there is* $\pi \in \mathfrak{M}$ *such that* $\pi \models \phi$ *and* $\pi \models \phi_i$*;*

(P4) for each $\phi \in \Phi$*, there is* $i \in \{1, \ldots, n\}$ *with* $\Phi_i \models \phi$*;*

(P5) there is no Φ' *such that* $\Phi' \models \Phi$ *and* $\Phi \not\models \Phi'$ *and* Φ' *satisfies (P1)-(P4).*

Considering the postulates in turn, we give an intuition behind the formalisation. The first postulate, P1, merely expresses that the statements in the middle ground should in itself make sense and be non-trivial. The second postulate, P2, expresses that whenever the stakeholders statements are not contradictory, the middle ground should simply consist of a union of their statements or a logical equivalent ($\equiv$). P3 expresses that any statement in the middle ground should be consistent with any individual statement of any of the stakeholders. Though, the middle ground might still oppose a collection of stakeholder preferences. P4 demands that any statement in the middle ground is justified by a stakeholder who's statements demand it. This is to prevent adding unnecessary statements to the middle ground. Finally, P5 ensures that among the sets of statements that satisfy P1-P4, the middle ground is maximal in the sense that it cannot be implied by another (non-equivalent) set.

To check that a middle ground is well defined, we need to consider the case of consistent stakeholders. It is easy to see that their joint statements satisfy the middle ground postulates.

Proposition 1. *If $\bigcup_{i=1}^{n} \Phi_i$ is consistent, then $\bigcup_{i=1}^{n} \Phi_i$ is a middle ground (Definition 2), that is, it satisfies P1-P5.*

3.2 Existence of Middle Ground

The satisfaction of P1-P4 is sufficient for the existence. For this, we note that the $\models$-relation is transitive, i.e., for $\Phi \models \Phi'$ and $\Phi' \models \Phi''$ we have $\Phi \models \Phi''$. Thus, $\models$ is acyclic for non-equivalent statements, i.e., there exists no chain of non-equivalent sets of statements $\Phi^1, \ldots, \Phi^k$ such that $\Phi^i \models \Phi^{i+1}$ for $i = 1, \ldots, k-1$ and $\Phi^k \models \Phi^1$. In consequence, since we assume $\mathcal{L}$ is finite, there exists a *dominating* set Φ such that there exists no other non-equivalent set $\Phi' \models \Phi$. Restricting $\models$ to sets of statements that satisfy P1-P4 preserves this. Using this observation and Proposition 1 the following holds.

Proposition 2. *Let $\Phi_1, \ldots, \Phi_n$ be non-trivial sets of statements. If there exists a set of statements Φ that satisfies P1, P3, and P4 then a middle ground exists for $\Phi_1, \ldots, \Phi_n$.*

Further, by using that $\pi \models \Phi$ implies $\pi \models \phi$ for $\phi \in \Phi$ and transitivity of $\models$, we can show that to check existence of a middle ground it is sufficient to consider *single* statements rather than *sets* of statements.

Proposition 3. *Let $\Phi_1, \ldots, \Phi_n$ be non-trivial sets of statements. If there exists a set of statements Φ that satisfies P3 and P4 for $\Phi_1, \ldots, \Phi_n$ then any statement φ such that $\phi \models \varphi$ for some $\phi \in \Phi$ satisfies P3 and P4.*

Similarly, one can also show that any φ with $\Phi \models \varphi$ satisfies P3 if Φ satisfies P3. The same is not true for P4.

However, if their union is consistent then we can show that it satisfies P1-P4 and thus, since they individually satisfy P5, they must be logically equivalent. Thus, middle grounds are either equivalent or inconsistent together.

Proposition 4. *Let Φ and Φ' be two sets of statements that are middle grounds for stakeholder statements $\Phi_1, \ldots, \Phi_n$. Then either $\Phi \equiv \Phi'$ or $\Phi \cup \Phi'$ is inconsistent.*

3.3 Construction of Middle Grounds

We can show that we can construct a middle ground with the help of the following algorithm, if there exists one. While not computationally efficient in general, this algorithm exploits the result of Proposition 3 by only considering satisfaction of P1, P3 and P4 for single statements rather than sets. This makes Algorithm 1 tractable for cases in which consistency and deduction problems are efficient.

Algorithm 1: Middle Ground for Statements

```
Input  : Non-trivial statement sets $\Phi_1, \ldots, \Phi_n \subseteq \mathcal{L}$
Output : Set of all middle grounds (up to equivalence).
1 if $\bigcup_{i=1}^{n} \Phi_i$ is consistent then return $\{\bigcup_{i=1}^{n} \Phi_i\}$ ;
2 $\Psi_1 := \{\varphi \in \mathcal{L} \mid \varphi \text{ non-trivial}\}$ ;
3 $\Psi_3 := \{\varphi \in \mathcal{L} \mid \forall \varphi' \in \bigcup_{i=1}^{n} \Phi_i \colon \{\varphi, \varphi'\} \text{ consistent}\}$ ;
4 $\Psi_4 := \{\varphi \in \mathcal{L} \mid \exists i \in \{1, \ldots, n\} \colon \Phi_i \models \varphi\}$ ;
5 return the set of all cardinality-maximal consistent
    subsets of $\Psi_1 \cap \Psi_3 \cap \Psi_4$ (possibly empty) ;
```

Theorem 5. *Algorithm 1 returns the (possible empty) set of all middle grounds (up to logical equivalence) for non-trivial sets of stakeholder statements.*

Proof Sketch. If $\bigcup_{i=1}^{n} \Phi_i$ is consistent, then by Line 1 the algorithm returns $\bigcup_{i=1}^{n} \Phi_i$ which, by P2 is the only middle ground (up to logical equivalence). Then, assume $\bigcup_{i=1}^{n} \Phi_i$ is inconsistent. In Lines 2-4, Algorithm 1 constructs the sets Ψ_i of $\phi \in \mathcal{L}$ that, individually, satisfies P_i, with $i \in \{1, 3, 4\}$. We show that Φ is a middle ground iff Φ is equivalent to a cardinality-maximal consistent subset of $\Psi := \Psi_1 \cap \Psi_3 \cap \Psi_4$, returned by Algorithm 1 (Line 5) (note that Ψ can be empty).

One can argue with the help of the postulates P1-4 and Proposition 3, that if Φ is a middle ground then Φ is equivalent to a consistent subset of Ψ. Next, one can show that any consistent subset Φ of Ψ satisfies P1-P4. Note that elements of a set of statements satisfy P3 and P4 individually, then the set also satisfies P3 and P4. We can now show that any cardinality-maximal subset Φ of Ψ that is consistent satisfies P5. Assume for contradiction that another set of statements Φ' satisfys P1-P4 and $\Phi' \models \Phi$ and $\Phi \not\models \Phi'$. One can now show $\Phi \cup \Phi'$ is inconsistent which implies $\Phi \not\models \Phi'$ — a contradiction.

We have shown that any middle ground is equivalent to a consistent subset of Ψ, and any cardinality-maximal consistent subset of Ψ is a middle ground. By Proposition 4, any two middle grounds are either equivalent or their union is inconsistent. Now, any non-cardinality-maximal consistent subset Φ of Ψ is consistent with one that is cardinality-maximal and thus is a middle ground. Hence, any middle ground is equivalent to a cardinality-maximal consistent subset of Ψ. □

4 Middle Grounds for Preferences

Here, we instantiate the general framework of Section 3 for the satisfaction system Λ with the language of preferences (Definition 6) and hierarchical models (Definition 3). We also consider the special case of Λ where we consider the class of lexicographic models (Definition 4). We start by formally defining all necessary notions and then analyse the complexity of deciding the existence of a middle ground.

4.1 Hierarchical Preferences

Variables and Alternatives: Let V be a set of m *variables* (or features) which describe alternatives. For each variable $v \in V$, let $\underline{v}$ denote its *domain*, i.e., the set of possible values of v. Assume that $\underline{v}$ is finite and contains more than one element. An *alternative* is an element of $\underline{V} = \prod_{v \in V} \underline{v}$ i.e., an assignment to all the variables. For alternative $\alpha \in \underline{V}$ and variable $v \in V$, let $\alpha(v) \in \underline{v}$ be the value α assigns to v.

Example 2 (cont.). *As before, we consider a setting similar to that in the Moral Machine Experiment [Awad* et al.*, 2018]. More concretely, let the alternatives be described by three variables with values between* 0 *and* 5 *as domains, such that* $\underline{V} = \underline{\mathsf{adult}} \times \underline{\mathsf{child}} \times \underline{\mathsf{dog}}$. *Consider the alternatives:*

$$\alpha = (1, 4, 0), \quad \beta = (2, 3, 3), \quad \gamma = (1, 3, 5).$$

Then, α *describes a set of 1 adult, 4 children, and 0 dogs. Similarly,* β *and* γ *specify sets of adults, children, and dogs.*

A hierarchical model consists of a hierarchy over variables. At each level of the hierarchy, we combine the variable assignments by a commutative and associative operator $\bigoplus$. Here, we assume that value domains of variables are compatible, i.e., there exists an operator $\bigoplus$ that can combine any subset of variables in a meaningful way, and there exists a natural order relation over the value domains as well as over values of combinations of variables. We can then compare alternatives by a lexicographic order. That is, we compare alternatives first based on the value combinations of the first-level variables; only if these are equal is the combination of the next most important variables considered, and so on.

Definition 3 (Hierarchical Model). *A* hierarchical model*, or simply model,* π *over variables* V*, is defined to be a non-empty sequence of the form* $(Y_1, \dots, Y_k)$*. Here* $Y_1, \dots, Y_k \subseteq V$ *are* k *non-empty sets of variables in* V.

Definition 4 (Lexicographic Model). *A* lexicographic model *is a hierarchical model with singleton variable sets. With an abuse of notation, we write such sequences as* $(v_1, \dots, v_k)$*, where* $v_1, \dots, v_k \in V$.

Our definitions are very similar to the models defined by Wilson *et al.* (2015), but differ in two points. First, we assume that neither hierarchical nor lexicographic models can be empty sequences. The corner case of empty models is a technical detail but, as becomes clearer in the following, does not contribute meaningful inference of preference statements. A more important difference is that, by our definition, hierarchical models may have non-disjoint sets of variables. That is, it would be possible to express that the number of humans is most important and the number of children is second most important, since children would appear in two levels of the importance order. Our definition is, in this latter point, a generalisation of the models defined in [Wilson *et al.*, 2015].

For any hierarchical models together with a commutative and associative operator $\bigoplus$ we define an order relation $\succeq_\pi$ over alternatives (omitting $\bigoplus$ for readability).

Definition 5 (Order Relation $\succeq_\pi$). *Let* V *be variables and* $\bigoplus$ *a commutative and associative operator on the variable domains. Assume that there exists a total order relation* $\geq$ *on the variable domains and on* $\bigoplus$*-combinations of variable values. For a model* $\pi = (Y_1, \dots, Y_k)$ *over variables* V *the binary relation* $\succeq_\pi$ *on* $\underline{V}$ *is defined as follows.*
For alternatives $\alpha, \beta \in \underline{V}$*, we have* $\alpha \succeq_\pi \beta$ *if and only if*

(i) for all $i = 1, \dots, k$, $\bigoplus_{y \in Y_i} \alpha(y) = \bigoplus_{y \in Y_i} \beta(y)$*, or*

(ii) there exists $i \in \{1, \dots, k\}$ *s.t.*

- $\bigoplus_{y \in Y_i} \alpha(y) > \bigoplus_{y \in Y_i} \beta(y)$ *and*
- $\bigoplus_{y \in Y_j} \alpha(y) = \bigoplus_{y \in Y_j} \beta(y)$ *for all* $j < i$.

The order relation $\succeq_\pi$ is a total pre-order on $\underline{V}$, i.e., reflexive, transitive and total. The order relation is not necessarily complete as it, e.g., does not necessarily include all variables. Thus, two alternatives might appear to be equivalent under $\succeq_\pi$ whereas they are different elements in $\underline{V}$.

The corresponding strict relation $\succ_\pi$ is given by $\alpha \succ_\pi \beta$ if and only if (ii) is satisfied, i.e., there exists $i \in \{1, \dots, k\}$ such that $\bigoplus_{y \in Y_i} \alpha(y) > \bigoplus_{y \in Y_i} \beta(y)$ and for all $j < i$, $\bigoplus_{y \in Y_j} \alpha(y) = \bigoplus_{y \in Y_j} \beta(y)$. The corresponding equivalence relation $\equiv_\pi$ is given by $\alpha \equiv_\pi \beta$ if and only if (i) is satisfied, i.e., for all $i = 1, \dots, k$, $\alpha(Y_i) = \beta(Y_i)$.

Example 3 (cont.). *Consider the alternatives in Example 2. If, it is desirable to save as many living beings as possible, then the natural order is "the more the better". As the domains are compatible (they are all the same) we could for example take the usual addition as the operator* $\bigoplus$. *Consider* $\pi = (\{\mathsf{adult}, \mathsf{child}\}, \{\mathsf{child}\})$. *This hierarchical model expresses that the number of humans to be saved from a car crash is the most important. Only if they are equal do we consider the number of children. Dogs are irrelevant in the comparison. Under the order relation induced by* π*, we have that* α *is strictly preferred to* γ *(and* β*),* $\alpha \succ_\pi \gamma$ *and thus* $\alpha \not\preceq_\pi \gamma$.

Definition 6 (Preference Language $\mathcal{L}$, [Wilson *et al.*, 2015]). *We define the language of non-strict and strict preference statements that are simple comparisons of alternatives* $\underline{V}$ *as:*

$$\mathcal{L} = \{\alpha \geq \beta \mid \alpha, \beta \in \underline{V}\} \cup \{\alpha > \beta \mid \alpha, \beta \in \underline{V}\}$$

We add parenthesis around preference statements when they appear in sequence, e.g. $(\alpha \geq \beta), (\gamma < \delta)$, for visualization. As Wilson *et al.* (2015), we define the meaning of these statements, and a satisfaction relation $\models$ between hierarchical models π and statements, in correspondence to $\succeq_\pi$.

Definition 7 (Satisfaction Relation $\models$). *Let* π *be a hierarchical model, and* $\alpha, \beta \in \underline{V}$ *alternatives.*

- *We say that* π *satisfies the non-strict statement* $\alpha \geq \beta$*, denoted by* $\pi \models \alpha \geq \beta$*, if and only if* $\alpha \succeq_\pi \beta$.
  That is, under π*,* α *is at least as preferred as* β.
- *We say that* π *satisfies the strict statement* $\alpha > \beta$*, denoted by* $\pi \models \alpha > \beta$*, if and only if* $\alpha \succ_\pi \beta$.
  That is, under π*,* α *is strictly preferred to* β.

Through $\mathcal{L}$, a stakeholder can express indifference between α and β via the statements $\alpha \geq \beta$ and $\beta \geq \alpha$ together. Further, as Wilson *et al.* (2015) already state, because $\succeq_\pi$ is a total pre-order over the alternatives, $\pi \not\models \alpha \geq \beta$ is equivalent to $\pi \models \beta > \alpha$. For this reason, we omit the definition of negated statements in $\mathcal{L}$. The notion of entailment is as in Definition 1.

Example 4 (cont.). *The statement* $\beta > \alpha$*, i.e.,* β *is strictly preferred to* α*, intuitively implies that any model* π *of the statement contains at least one set of individuals (that are important to the stakeholder) from which there are strictly more beings saved in* β *than in* α*, e.g.,* $(\{\mathsf{adult}, \mathsf{child}\}, \{\mathsf{dog}\}) \models \beta > \alpha$. *While* $(\{\mathsf{adult}, \mathsf{child}\}, \{\mathsf{dog}\}) \models \alpha > \gamma$*, we cannot deduce* $\alpha > \gamma$ *from* $\beta > \alpha$ *(*$\{\beta > \alpha\} \not\models \gamma > \alpha$*) because also* $(\{\mathsf{dog}\}) \models \beta > \alpha$ *and* $(\{\mathsf{dog}\}) \not\models \alpha > \gamma$.

4.2 Non-Uniqueness and Non-Existence

As a first result, we note that there may be more than one middle ground for preference statements in $\mathcal{L}$.

Theorem 6. *There exist sets of stakeholder statements in $\mathcal{L}$ that admit multiple non-equivalent middle grounds.*

Proof Sketch. Consider the following alternatives defined over four binary variables $V = \{x, y, z, w\}$:

	x	y	z	w
$\alpha =$	1	0	0	0
$\beta =$	0	1	0	0
$\alpha' =$	0	0	1	0
$\beta' =$	0	0	0	1
$\gamma =$	1	0	1	0
$\delta =$	0	1	0	1

For simplicity, we assume that the value of any $\oplus$-combination of variables is the same for all alternatives and omit such values in the table on the left.

Consider two stakeholders expressing non-trivial statements:

$$\Phi_1 = \{(\alpha > \beta), (\alpha' > \beta')\}, \quad \Phi_2 = \{(\beta > \alpha), (\beta' > \alpha')\}.$$

The stakeholder's statements are consistent individually, but inconsistent together. Thus, the union of Φ_1 and Φ_2 cannot be a middle ground. One can show that there are at least two non-equivalent middle grounds for Φ_1 and Φ_2 with help of the two statements $\psi_1 = \gamma > \delta$ and $\psi_2 = \delta > \gamma$. In particular:

1. ψ_1 and ψ_2 are individually non-trivial;
2. ψ_1 and ψ_2 are inconsistent together;
3. for all $i, j \in \{1, 2\}$ and all $\phi_i \in \Phi_i$, there is π such that $\pi \models \psi_j$ and $\pi \models \phi_i$;
4. for $i \in \{1, 2\}$, $\Phi_i \models \psi_i$.

To conclude, we claim that there are at least two non-equivalent middle grounds: one that contains ψ_1 and another one that contains ψ_2. Indeed, Points (1), (3), (4) and Theorem 5 imply that that there is a middle ground for Φ_1 and Φ_2 that contains ψ_1 (plus possibly other statements, so as to satisfy P5) and a middle ground for Φ_1 and Φ_2 that contains ψ_2. Point (2) implies that there is no middle ground that contains both ψ_1 and ψ_2 (otherwise P1 would be violated). So there are two non-equivalent middle grounds for Φ_1 and Φ_2. □

Further, we show that a middle ground may not exist.

Theorem 7. *There exist sets of stakeholder statements in $\mathcal{L}$ that admit no middle ground.*

Proof. Consider alternatives defined over two binary variables $V = \{x, y\}$, and an operator $\oplus$ that resembles the logical $\wedge$:

	x	y	$x \oplus y$
$\alpha =$	1	0	0
$\beta =$	0	1	0
$\gamma =$	1	1	1
$\delta =$	0	0	0

By the convention $1 > 0$, any hierarchical model entails $\alpha \geq \delta$, $\beta \geq \delta$ and $\gamma > \delta$, as well as $\gamma \geq \alpha$ and $\gamma \geq \beta$. Further, no model satisfies $\delta > \gamma$.

The set of non-trivial statements in this case is given by $\mathcal{N} = \{(\gamma > \alpha), (\gamma \leq \alpha), (\gamma > \beta), (\gamma \leq \beta), (\alpha > \beta), (\alpha \geq \beta), (\alpha < \beta), (\alpha \leq \beta), (\alpha > \delta), (\alpha \leq \delta), (\beta > \delta), (\beta \leq \delta)\}$.

Consider two stakeholders with preference statements:

$$\Phi_1 = \{\alpha \geq \gamma\} \quad \text{and} \quad \Phi_2 = \{\beta \geq \gamma\}.$$

These statements are consistent individually, but inconsistent together. In particular, the only hierarchical model that satisfies Φ_1 is $(\{x\})$. Thus Φ_1 entails non-trivial statements $\{(\gamma > \beta), (\alpha \geq \gamma), (\alpha > \beta), (\alpha \geq \beta), (\alpha > \delta), (\delta \geq \beta)\} \subseteq \mathcal{N}$. None of these statements is consistent with Φ_2. By symmetry, the only hierarchical model satisfying Φ_2 is $(\{y\})$ and none of the entailed statements from Φ_2 are consistent with Φ_1.

By P4 any statement in the middle ground is entailed by some stakeholders statements. However, by P3 and because the stakeholders have only one statement each, the middle ground needs to be consistent with the stakeholders statements. As argued above, there is no such middle ground. □

4.3 Deciding Existence of a Middle Ground

Through Proposition 2 we have established that for the existence of a middle ground it is sufficient to check whether there exists a set of statements that satisfies P1, P3, and P4. Further, we found that, by Proposition 3, it is sufficient to only check for the existence of single (non-trivial) statements that satisfy P3 and P4. For preference statements of language $\mathcal{L}$ we can further narrow down which statements shall be investigated to determine existence of a middle ground.

As a consequence of Proposition 3, and because $(\alpha > \beta) \models (\alpha \geq \beta)$, we have the following relation between strict and non-strict statements satisfying P3 and P4.

Corollary 8. *Let $\Phi_1, \ldots, \Phi_n \subseteq \mathcal{L}$ be non-trivial sets of statements. If the strict statement $\alpha > \beta$ satisfies P3 and P4 then its non-strict version $\alpha \geq \beta$ satisfies P3 and P4.*

Here the non-strict version, while satisfying P3 and P4, might be trivial (i.e., violating P1) even if the strict statement is non-trivial. However, we can observe that this can only happen in a specific case.

Lemma 9. *If $\alpha > \beta$ is non-trivial then either $\alpha \geq \beta$ is non-trivial or*

- *$\bigoplus_{y \in Y} \alpha(y) \geq \bigoplus_{y \in Y} \beta(y)$ for all $Y \subseteq V$, and*
- *there exists $Y \subseteq V$ with $\bigoplus_{y \in Y} \alpha(y) = \bigoplus_{y \in Y} \beta(y)$.*

Thus, if a strict statement is non-trivial, its "non-strict version" cannot be a contradiction. Further, it can only be a tautology, if there is a variable set that is indifferent.

The discussion above together with Propositions 2 and 3 allows us now to specify sets of non-trivial strict and non-strict statements that are sufficient to check w.r.t. P3 and P4 to guarantee the existence of a middle ground.

Corollary 10. *Let $\Phi_1, \ldots, \Phi_n$ be non-trivial sets of statements. There exists a middle ground that includes a strict or a non-strict statement if and only if one of the following statements satisfies P3 and P4:*

$$\{\alpha \geq \beta \mid \alpha, \beta \in \underline{V} \text{ s.th. } \alpha \geq \beta \text{ non-trivial}\}$$
$$\cup \{\alpha > \beta \mid \alpha, \beta \in \underline{V} \text{ s.th. } \alpha > \beta \text{ non-trivial}, \alpha \geq \beta \text{ trivial}\}.$$

While we can exactly determine the sets of statements in Corollary 10 and they are finite, they are exponentially large on the number of variables. As we see next, checking the postulates of the definition of a middle ground is not in P for hierarchical models, unless P=NP and P=coNP. Thus, we consider lexicographic models, which are a special case of hierarchical models. For these we can narrow down the set of preference statements that need to be checked for satisfying P3 and P4 to only $2 \cdot |V|$ statements. Since checking the satisfaction of P3 and P4 is polynomial for lexicographic models [Wilson *et al.*, 2015], the existence of a middle ground is decidable in polynomial time.

Hierarchical Models

To analyse the complexity of deciding the existence of a middle ground for hierarchical models, we first show that it is NP-complete to decide consistency for hierarchical models.

Wilson *et al.* (2015) show that deciding $\Gamma \models \alpha \geq \beta$ is coNP-complete for their definition of hierarchical models which they call HCLP models, even if Γ is a set of non-strict statements [Wilson *et al.*, 2015]. Consequentially, deciding consistency of a set of statements Γ is NP-complete under HCLP models. However, as outlined before, HCLP models are slightly differently defined than hierarchical models and the construction of the reduction from 3SAT to prove their central result is not transferable to hierarchical models. In particular their Lemma 2 does not hold if variable sets in models are allowed to be non-disjoint.

To show NP-completeness of deciding consistency of a set of statements w.r.t. our definition of hierarchical models, we instead use a reduction from the Subset Sum Problem.[1] An instance of Subset Sum consists of a multi-set of integers $\mathcal{S}$ and a target integer T. The task is to decide whether there exists a multi-set $A \subseteq \mathcal{S}$ such that the sum of its element is T, i.e., $\sum_{a \in A} a = T$. This problem is NP-complete even if all integers in $\mathcal{S}$ are positive [Kleinberg and Tardos, 2006].

Theorem 11. *Deciding consistency of a set of preference statements is NP-complete w.r.t. hierarchical models with operators $\oplus$ that can be computed in time polynomial in the number of variables.*

Proof. To see that the consistency problem is in NP, we show that one can check in time polynomial in the number of variables and statements, whether a given hierarchical model $\pi = (Y_1, \ldots, Y_k)$ satisfies a set of given preference statements $\Gamma \subseteq \mathcal{L}$. That is, for every non-strict statement $(\alpha \succeq \beta) \in \Gamma$ we need to check whether $\bigoplus_{y \in Y_i} \alpha(y) \geq \bigoplus_{y \in Y_i} \beta(y)$ for all $i = 1, \ldots, k$, and, for every strict statement $(\alpha \succ \beta) \in \Gamma$, we need to additionally check whether there exists $i \in \{1, \ldots, k\}$ with $\bigoplus_{y \in Y_i} \alpha(y) \geq \bigoplus_{y \in Y_i} \beta(y)$. By our assumption on $\oplus$ this can be computed in polynomial time.

We show the completeness of the problem by a reduction from Subset Sum with positive integers. For this, let $\mathcal{S}$ be a multiset of positive integers and $T \in \mathbb{N}$ a target. We construct three preference statements that, when satisfied together, force a hierarchical model to contain a variable set that corresponds to a solution of Subset Sum for $\mathcal{S}$ and T.

Variables: We construct a variable v_a for each $a \in \mathcal{S}$ and one variable v_T. Denote the set of variables by V. Then $|V| = |\mathcal{S}| + 1$. Let the domain of each variable be $\mathbb{N}$.
Operator: The operator $\oplus$ is the normal addition.
Preference Statements: Consider preference statements

$$\Phi = \{(\alpha_T > \beta_T), (\alpha_\Sigma \geq \beta_\Sigma), (\beta_\Sigma \geq \alpha_\Sigma)\}$$

for alternatives $\alpha_T, \beta_T, \alpha_\Sigma, \beta_\Sigma$ that are defined as follows:

$$\alpha_T(v_c) = \begin{cases} 0 \text{ if } c \in \mathcal{S} \\ 1 \text{ if } c = T \end{cases} \qquad \beta_T(v) = 0 \ \forall v \in V$$

$$\text{and} \quad \alpha_\Sigma(v_c) = \begin{cases} c \text{ if } c \in \mathcal{S} \\ 0 \text{ if } c = T \end{cases} \qquad \beta_\Sigma(v_c) = \begin{cases} 0 \text{ if } c \in \mathcal{S} \\ T \text{ if } c = T. \end{cases}$$

Satisfaction: We first show that if there exists a hierarchical model satisfying Φ then there exists a Subset Sum solution. Then we show the reverse.

Assume that there is a hierarchical model π with $\pi \models \Phi$. Because $\alpha_T > \beta_T$ is a strict statement, but all variables are indifferent under α_T and β_T except v_T, π must contain v_T in some variable set. Let C be the first such variable set in π with $v_T \in C$. Then by $\pi \models (\alpha_\Sigma \geq \beta_\Sigma)$, either (1) C is preceded by another variable set C' in π, or (2) C contains other variables and $\sum_{v_c \in C} \alpha_\Sigma(v_c) \geq \sum_{v_c \in C} \beta_\Sigma(v_c) = T$. By $\pi \models (\beta_\Sigma \geq \alpha_\Sigma)$ and because integers in $\mathcal{S}$ are positive, case (1) is not possible. Thus assume that case (2) holds and let $A = \{a \in \mathcal{S} \mid v_a \in C \setminus \{v_T\}\}$. Then $\sum_{a \in A} a = \sum_{v_c \in C \setminus \{v_T\}} \alpha_\Sigma(v_c) \geq T$. Further, by $\pi \models (\beta_\Sigma \geq \alpha_\Sigma)$ and because there is no other variable set preceding C in π, we have $T = \sum_{v_c \in C} \beta_\Sigma(v_c) \geq \sum_{v_c \in C} \alpha_\Sigma(v_c) = \sum_{a \in A} a$. So, if $\pi \models \Phi$ then π contains a set of variables C corresponding to T and integers $A \subseteq \mathcal{S}$ such that $T = \sum_{a \in A} a$.

For the reverse, assume there exists a multiset A that is a subset of integers $\mathcal{S}$ with $T = \sum_{a \in A} a$. Then, as shown before, the hierarchical model $\pi = (\{v_a \mid a \in A\} \cup \{v_T\})$ satisfies Φ. Thus there exists a hierarchical model satisfying Φ iff there exists a subset of $\mathcal{S}$ that sums to T. Because construction of Φ is polynomial in the size of the Subset Sum instance, and Subset Sum is NP-complete, so is deciding consistency for hierarchical models and preference statements. □

Recall that for preference statements $\Gamma, \alpha > \beta$, we have $\Gamma \models (\alpha > \beta)$ if and only if $\Gamma \cup \{\alpha \leq \beta\}$ is inconsistent [Wilson *et al.*, 2017]. Similarly, $\Gamma \models (\alpha \geq \beta)$ if and only if $\Gamma \cup (\alpha < \beta)$ is inconsistent. Then Theorem 11 has the following consequences for the complexity of deduction and testing middle ground.

Corollary 12. *Deciding $\Gamma \models \varphi$ for preference statements Γ and φ w.r.t. hierarchical models is coNP-complete.*

Corollary 13. *Let $\Phi_1, \ldots, \Phi_n$ be non-trivial sets of statements. Let Φ be a given set of statements and consider the satisfaction relation w.r.t. hierarchical models. Then, deciding whether Φ satisfies (P1) is NP-complete and deciding whether Φ satisfies (P4) for $\Phi_1, \ldots, \Phi_n$ is coNP-complete.*

Lexicographic Models

For lexicographic models, we are able to decrease the set of statements in Corollary 10 that need to be considered to determine existence of a middle ground to a polynomial size. We

[1] The same proof can also be used to show NP-completeness for HCLP models with an addition for the (trivial) case of the empty HCLP model. It thus offers a more concise alternative to the proof of Wilson *et al.* (2017). More generally, it shows that deciding consistency is NP-complete even for models containing only one set of variables, and only three preference statements.

first show that it is sufficient to consider binary variables. Informally, this follows from the fact that lexicographic models only consider ordinal relations between the variable assignments and not their actual values.

Proposition 14. *Given statement $\phi = (\alpha \geq \beta)$, let $\phi^b = (\alpha^b \geq \beta^b)$ be the result of replacing $\alpha(v)$ and $\beta(v)$ in ϕ by*

- 1 *and* 0*, respectively, if* $\alpha(v) > \beta(v)$*;*
- 0 *and* 1*, respectively, if* $\alpha(v) < \beta(v)$*; and*
- 0 *and* 0*, respectively, if* $\alpha(v) = \beta(v)$*, for all* $v \in V$*.*

Under the assumption that $1 > 0$, we have for all lexicographic models π, that $\pi \models \phi$ iff $\pi \models \phi^b$.

We can further decrease the set of statements to consider for testing existence of a middle ground as follows.

Theorem 15. *Consider inference based on lexicographic models on variables V. Let $\vec{0}$ denote the vector of $|V|$ zeros and let $\vec{0}_v$ be the same as $\vec{0}$ but with 1 at the position corresponding to a variable $v \in V$. Similarly, let $\vec{1}$ denote the vector of $|V|$ ones and let $\vec{1}_v = \vec{1} - \vec{0}_v$. There exists a middle ground for non-trivial sets of statements $\Phi_1, \ldots, \Phi_n$ if and only if one of the following statements satisfy P3 and P4:*

$$\{\vec{1}_v \geq \vec{0}_v \mid v \in V\} \cup \{\vec{1}_v > \vec{0} \mid v \in V\}.$$

Proof. Suppose there exists a middle ground for $\Phi_1, \ldots, \Phi_n$ that includes a non-trivial statement ϕ. By Proposition 14, and under the assumption that $1 > 0$ we have for all lexicographic models π, that $\pi \models \phi$ iff $\pi \models \phi^b$. Here ϕ^b is the statement over binary variables as defined in Proposition 14. We can thus focus our following argumentation on ϕ^b instead of ϕ.

Assume that ϕ^b is a non-strict statement $\alpha \geq \beta$. Then because it is non-trivial, and in particular not a tautology, there exists variable v such that $\alpha(v) < \beta(v)$. Any lexicographic model that satisfies $\alpha \geq \beta$ must include another variable v' preceding v or not include v at all. Thus, any such model also satisfies $\vec{1}_v \geq \vec{0}_v$, i.e., $(\alpha \geq \beta) \models (\vec{1}_v \geq \vec{0}_v)$.

Now assume that ϕ^b is a strict statement $\alpha > \beta$. Then because it is non-trivial, by Lemma 9, it entails either (1) a non-trivial non-strict statement or (2) there exists a variable v such that $\alpha(v) = \beta(v) (= 0)$ and $\alpha(v') \geq \beta(v')$ for all variables $v' \in V \setminus \{v\}$. In case (1), by our arguments above, $\alpha > \beta$ also entails a non-trivial non-strict statement $\vec{1}_v \geq \vec{0}_v$ for some $v \in V$. For case (2), we show that $\alpha > \beta$ entails the statement $\vec{1}_v > \vec{0}$. Because $\alpha > \beta$ is a strict statement and because $\alpha(v) = \beta(v)$ the lexicographic model that *only* includes v does not satisfy $\alpha > \beta$. Thus, any lexicographic model satisfying $\alpha > \beta$ must also include some other variable v'. Any such model then also satisfies $\vec{1}_v > \vec{0}$.

As shown above, any middle ground of statements entails a statement in $\{\vec{1}_v \geq \vec{0}_v \mid v \in V\} \cup \{\vec{1}_v > \vec{0} \mid v \in V\}$. By Proposition 3, this statement satisfies P3 and P4.

For the converse direction, assume there exists a statement in $\{\vec{1}_v \geq \vec{0}_v \mid v \in V\} \cup \{\vec{1}_v > \vec{0} \mid v \in V\}$ that satisfies P3 and P4. Because all such statements are by construction non-trivial, they also satisfy P1. Then by Proposition 2 there exists a middle ground. □

Algorithm 2: Existence of Middle Ground

```
Input  : Sets of non-trivial preferences $\Phi_1, \ldots, \Phi_n$.
Output : 'yes' if it exists, 'no' otherwise.
if $\bigcup_{i=1}^n \Phi_i$ is consistent then return $\bigcup_{i=1}^n \Phi_i$ ;
$\mathcal{L}_b := \{(\vec{1}_v \geq \vec{0}_v), (\vec{1}_v > \vec{0}) \mid v \in V\}$;
$\Psi_3 := \{\varphi \in \mathcal{L}_b \mid \forall \psi \in \bigcup_{i=1}^n \Phi_i \colon \{\varphi, \psi\}$ consistent$\}$;
$\Psi_4 := \{\varphi \in \mathcal{L}_b \mid \exists i \in \{1, \ldots, n\} \colon \Phi_i \models \varphi\}$;
if $\Psi_3 \cap \Psi_4 = \emptyset$ then return 'no';
else return 'yes';
```

Wilson *et al.* (2015) establish that checking consistency or inference is polynomial-time solvable for lexicographic models and strict and non-strict preference statements. In consequence, we can check in polynomial-time whether a statement satisfies P3 or P4. By Theorem 15, we only need to do this for $2 \cdot |V|$ many statements for lexicographic models.

Corollary 16. *Let $\Phi_1, \ldots, \Phi_n$ be non-trivial sets of statements. Checking whether there exists a middle ground w.r.t. lexicographic models that includes a (non-trivial) strict or non-strict statement is polynomial-time solvable.*

We summarise the algorithm to decide existence of a middle ground in Algorithm 2. Here, the set of statements $\mathcal{L}_b$ has cardinality $2 \cdot |V|$. To construct Ψ_3 we need to perform $2 \cdot |V| \cdot |\bigcup_{i=1}^n \Phi_i|$ consistency checks between two statements. To construct Ψ_4 we need to perform $2 \cdot |V| \cdot n$ checks to see if a statement can be deduced from a stakeholder's statements.

One can also employ Algorithm 1 to *construct* a middle ground for lexicographic models. In case the stakeholders statements are inconsistent (which can be checked in polynomial time), by Proposition 14, one can then focus on binary statements instead of the complete language $\mathcal{L}$. The number of strict and non-strict statements over $|V|$ variables with binary domains is $O(2^{2|V|-1})$. We then need to check every one of the $O(2^{2|V|-1})$ statements on whether they satisfy P1, P3 and P4. Thus, while each of these tests individually can be done in polynomial time, Algorithm 1 remains exponential even for lexicographic models. It remains an open question whether there exist tractable algorithms to solve this problem.

5 Conclusion

We investigate the notion of middle ground, exploring its properties, including the fact that it may not exist or be unique. We establish necessary conditions for its existence and describe an algorithmic procedure for both existence checking and construction. Our case study focuses on preference statements, with a notable application in moral preferences and hiring: while the general problem is coNP-complete, we show that deciding existence is tractable for lexicographic models.

Our postulate P3 concerns statements in the middle ground individually and not the whole set. It remains open whether a stronger version of this or other postulates lead to more tractable algorithms or a unique middle ground. Other future work may analyse the tractability of constructing middle grounds for lexicographic models and explore the concept for non-preference-based languages, such as propositional logic.

Acknowledgments

This work was supported by the Research Council of Norway through its Centre of Excellence: Integreat - The Norwegian Centre for knowledge-driven machine learning, project number 332645.

References

[Adler, 2016] Matthew D. Adler. Aggregating moral preferences. *Economics and Philosophy*, 32(2):283–321, 2016.

[Aiguier *et al.*, 2018] Marc Aiguier, Jamal Atif, Isabelle Bloch, and Céline Hudelot. Belief revision, minimal change and relaxation: A general framework based on satisfaction systems, and applications to description logics. *Artif. Intell.*, 256:160–180, 2018.

[Awad *et al.*, 2018] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563, 11 2018.

[Bienvenu *et al.*, 2010] Meghyn Bienvenu, Jérôme Lang, and Nic Wilson. From preference logics to preference languages, and back. In *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, 2010.

[Botan *et al.*, 2023] Sirin Botan, Ronald de Haan, Marija Slavkovik, and Zoi Terzopoulou. Egalitarian judgment aggregation. *Auton. Agents Multi Agent Syst.*, 37(1):16, 2023.

[Boutilier *et al.*, 2004] Craig Boutilier, Ronen I Brafman, Carmel Domshlak, Holger H Hoos, and David Poole. Cp-nets: A tool for representing and reasoning withconditional ceteris paribus preference statements. *Journal of artificial intelligence research*, 21:135–191, 2004.

[Delgrande and Schaub, 1997] J. P. Delgrande and T. Schaub. Compiling reasoning with and about preferences into default logic. In *IJCAI*, pages 168–175. Morgan Kaufmann, 1997.

[Delgrande *et al.*, 2018] James P. Delgrande, Pavlos Peppas, and Stefan Woltran. General belief revision. *J. ACM*, 65(5):29:1–29:34, 2018.

[Freedman *et al.*, 2020] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

[Gärdenfors, 1986] P. Gärdenfors. Belief Revisions and the Ramsey Test for Conditionals. *The Philosophical Review*, 95(1):81–93, 1986. http://www.jstor.org/stable/2185133.

[George and Wilson, 2016] Anne-Marie George and Nic Wilson. Preference inference based on pareto models. In *Scalable Uncertainty Management: 10th International Conference, SUM 2016, Nice, France, September 21-23, 2016, Proceedings 10*, pages 170–183. Springer, 2016.

[Guimarães *et al.*, 2023] Ricardo Guimarães, Ana Ozaki, and Jandson S. Ribeiro. Finite based contraction and expansion via models. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *AAAI*, pages 6389–6397. AAAI Press, 2023.

[Horty, 1994] John F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23(1):35–65, 1994.

[Ju *et al.*, 2020] F. Ju, K. Nygren, and T. Xu. Modeling legal conflict resolution based on dynamic logic. *Journal of Logic and Computation*, 31(4):1102–1128, 2020.

[Kleinberg and Tardos, 2006] Jon Kleinberg and Eva Tardos. *Algorithm design*. Pearson Education India, 2006.

[Kollingbaum *et al.*, 2008] Martin J. Kollingbaum, Wamberto W. Vasconcelos, Andres García-Camino, and Tim J. Norman. Managing conflict resolution in norm-regulated environments. In Alexander Artikis, Gregory M. P. O'Hare, Kostas Stathis, and George Vouros, editors, *Engineering Societies in the Agents World VIII*, pages 55–71. Springer Berlin Heidelberg, 2008.

[Konieczny and Pérez, 2011] Sébastien Konieczny and Ramón Pino Pérez. Logic based merging. *J. Philosophical Logic*, 40(2):239–270, 2011.

[Liao *et al.*, 2023] Beishui Liao, Pere Pardo, Marija Slavkovik, and Leendert van der Torre. The Jiminy Advisor: Moral Agreements among Stakeholders Based on Norms and Argumentation. *Journal of Artificial Intelligence Research*, 77:737–792, 2023.

[Moodley *et al.*, 2011] Kodylan Moodley, Thomas Meyer, and Ivan José Varzinczak. Root justifications for ontology repair. In Sebastian Rudolph and Claudio Gutierrez, editors, *RR*, volume 6902 of *Lecture Notes in Computer Science*, pages 275–280. Springer, 2011.

[Ozaki *et al.*, 2024] Ana Ozaki, Anum Rehman, and Marija Slavkovik. Finding middle grounds for incoherent horn expressions: the moral machine case. *Auton. Agents Multi Agent Syst.*, 38(2):50, 2024.

[Wilson and Montazery, 2016] Nic Wilson and Mojtaba Montazery. Preference inference through rescaling preference learning. International Joint Conferences on Artificial Intelligence Organization, 2016.

[Wilson *et al.*, 2015] Nic Wilson, Anne-Marie George, and Barry O'Sullivan. Computation and complexity of preference inference based on hierarchical models. In *24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3271–3277. AAAI Press, 2015.

[Wilson *et al.*, 2017] Nic Wilson, Anne-Marie George, and Barry O'Sullivan. Preference inference based on hierarchical and simple lexicographic models. *Journal of Applied Logics - IfCoLog Journal*, 4 (7):1997–2038, 2017.