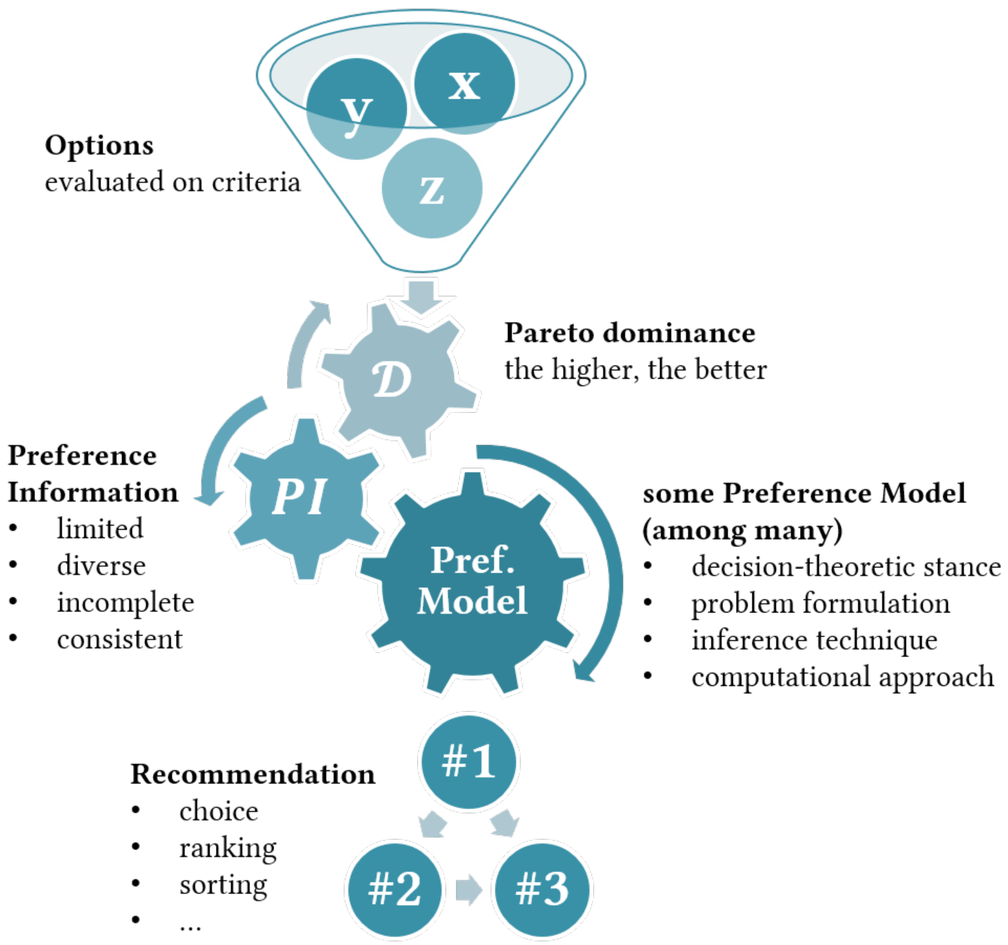
Challenges in Explaining Decisions (and some links to Social Choice)

Khaled Belahcene[1], C. Labreuche[2], N. Maudet[3], V. Mousseau[1] and W. Ouerdane[1]

■ Multi-Criteria Decision Aiding (MCDA)



From interpretability...

- Axiomatized MCDA models claim "interpretability", but they are hardly intelligible by themselves;
- MCDA is structurally a "human in the loop" methodology and process. The Decision Maker's grasp of the stakes is crucial;
- Existing explanation frameworks, designed to complement Decision Support Systems non-specifically MCDA, are too lightweight.

■ Landmarks

1. I. Alvarez : *Explaining the result of a Decision Tree to the End-User.* ECAI 2004: 411-415
   ☞ An explanation is more than an arbitrary trace of the decision process
2. S. Greco, V. Mousseau, and R. Slowinski. *Ordinal regression revisited: multiple criteria ranking using a set of additive value functions.* European Journal of Operational Research, 191(2):415-435, 2008.
   ☞ Necessary and possible preference statements in an additive utility MCDA model
3. W. Ouerdane, N. Maudet, A. Tsoukiàs: *Argument Schemes and Critical Questions for Decision Aiding Process.* COMMA 2008: 285-296
   ☞ An interactive MCDA framework based on argumentation techniques. Model selection is addressed from a user-centric perspective.
4. C. Labreuche: *A general framework for explaining the results of a multi-attribute preference model.* Artif. Intell. 175(7-8): 1410-1448 (2011)
   ☞ A principled way of selecting arguments supporting decisions in MCDA models assigning weights to criteria
5. C. Labreuche, N. Maudet, W. Ouerdane: *Justifying Dominating Options when Preferential Information is Incomplete.* ECAI 2012: 486-491
   ☞ An effective engine explaining necessary preference statements in a weighted Condorcet model, with duality techniques
6. Olivier Cailloux and Ulle Endriss, *Arguing about Voting Rules*, AAMAS-2016
   ☞ Voting rules are promoted by exhibiting meaningful situations showcasing the underlying axioms
7. Spiegler, R. *Equilibrium in Justifiable Strategies: A Model of Reason-Based Choice in Extensive-Form Games.* The Review of Economic Studies, 69(3), 2002
   ☞ In a strategic context, players are restricted to strategies they can account for, leading to a new definition of equilibrium

Contributions

■ Explaining robust additive utility models by sequences of preference swaps, *Theory and Decision*, in print

➤ Problem statement

PI : a set of pairwise preference statements

Model : any satisfying Pareto, Transitivity and Cancellation axioms, e.g.

- any particular Additive Value model, i.e. $x \succsim y \iff \sum V_i(x_i) \geq \sum V_i(y_i)$
- x is necessarily preferred to y iff $V(x) \geq V(y)$ for every possible Additive Value model correctly representing the PI.

Recommendation : a preference statement $x \succsim y$

➤ Proposed explanation

A sequence of options $x = e_0 \succsim e_1 \succsim \cdots \succsim e_{n-1} \succsim e_n = y$

- establishing the preference of x over y (transitivity)
- two adjacent options differ only on 1 (dominance) or 2 (trade-off) criteria

➤ Results and Challenges

Existence ? Bound on sequence length ? Computation ?

- **Necessary Preference + binary PI :** Explanations have a term-by-term structure. Efficient algorithm for existence and actual computation. Explanations can be kept short. Proofs use PL/duality, and graph/flows techniques.
- **general case :** Open issues. We provide an example where there is no upper bound on the length of the shortest possible explanations

■ Accountable Classification without frontiers, DA2PL'16, submitted

➤ Design principles favoring *Accountability*

No jargon. No values. No frontiers. No compensation. No inference.

- an object can not outrank any object assigned to a strictly better class;
- an object outranks objects assigned to a strictly worse class;

➤ Implementation

- the model **observes** every pair of reference objects not assigned to the same class
- it **learns** sets of *sufficient*, *insufficient*, or undecided coalitions of criteria, accounting for monotonicity
- for a given candidate, it **recommends** every *possible* assignment not contradicting its principles
- it **explains** its recommendation with supporting statements instantiating specified *argument schemes*

Object	a	b	c	d	*Assignment* ⟶
A_1	A	A	2.5	*False*	★★★
A_2	A	B	2.1	*True*	★★★
B_1	B	B	1.3	*True*	★★
B_2	A	C	3.7	*False*	★★
C_1	B	C	1.6	*True*	★
C_2	C	C	4.1	*False*	★
Z_1	B	B	1.1	*False*	?
Z_2	B	A	1.8	*False*	??
Z_3	A	B	1.2	*False*	???

	★★★		★★		★		?	?	?
	A_1	A_2	B_1	B_2	C_1	C_2	Z_1	Z_2	Z_3
A_1			abc	abd	abc	abd	**abcd**	**abcd**	**abcd**
A_2			abcd	abd	abc	abd	**abcd**	acd	**abcd**
B_1	d	bd			abd	abd	**abcd**	ad	bcd
B_2	acd	ac			abc	abd	cd	acd	acd
C_1	d	d	acd	bd			acd	acd	cd
C_2	cd	c	c	bcd			cd	cd	cd
Z_1	d	b	(ab)	bd	(ab)	**abd**			
Z_2	bd	bc	**abc**	bd	(ab)	**abd**			
Z_3	(ab)	(ab)	(ab)	**abd**	(ab)	**abd**			

For example Z_2 should at least be assigned ★★, as Z_2 is at least as good as B_1 on every criteria except d, *and* abc *is established as sufficient by the comparison A_1 vs C_1.*

... to Accountability.

- Accountability is the ability of a human decision maker to own a recommendation made by the system and to *transfer* this ownership
- It suits MCDA better than mere trust, transparency, or persuasiveness, and leads to actual implementation
- Explanations require in-depth understanding of the preference models, and pose interesting computational challenges
- It mixes Decision Theory, Optimization techniques, and several Artificial Intelligence approaches (e.g. knowledge representation, argumentation)

Connections to Computational Choice

- **Structure** : **MCDA and CSC are structurally close**, as Choice and Ranking mirrors **Voting**, and Ordinal Sorting mirrors **Judgment Aggregation**
- **Techniques** : Explaining the **result** of a Social Choice algorithm, or the **selection of a particular procedure**, could borrow techniques and insights
- **Applications** : Accountability is particularly needed in situations addressed simultaneously by MCDA and CSC, such as committee decisions
- **Complexity** : designing a model behaving well w.r.t. Accountability, incorporating requirements for accountability in adversarial contexts, modelling the collective reconstruction of explanations in a context similar to gossip,...

(1) LGI, CentraleSupélec, Univ. Paris Saclay , (2) Thales, (3) LIP6, Univ. Pierre et Marie Curie