# Interactive Explanation and Elicitation for Multiple Criteria Decision Analysis

**Vincent Mousseau & Wassila Ouerdane**

Laboratoire Génie Industriel

In collaboration with : Kh. Belahcene (LGI), Ch. Labreuche (Thales) and N. Maudet (LIP6)

CentraleSupélec

IRT SystemX–April 11, 2018

# Contents

Motivations

Introduction to Multiple Criteria Decision Aiding
- Basic MCDA concepts
- Preference Elicitation

Explanation schemes in MCDA context
- Pairwise comparisons
- Ordinal Sorting

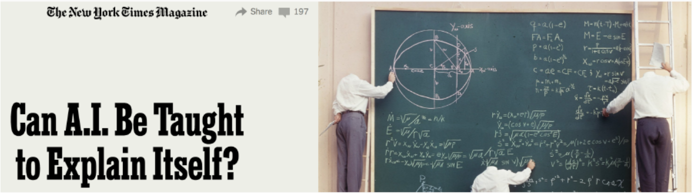
Future prospects and applications

# Motivations

- new regulations (eg. GDPR)
- raising concern in the society : making A.I. systems trustable !

Featured in mainstream press, related to prominent applications :

- automated decisions for autonomous vehicles
- loan agreements
- Admission Post Bac/ParcourSup

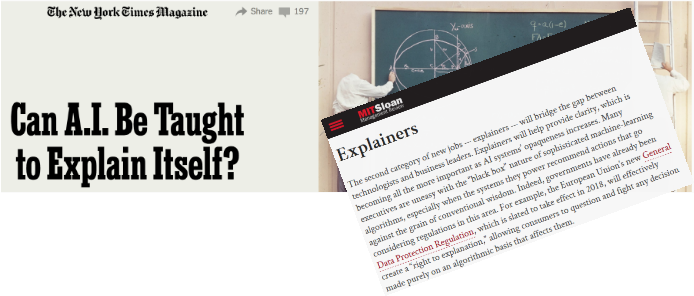


# Motivations

- new regulations (eg. GDPR)
- raising concern in the society : making A.I. systems trustable !

Featured in mainstream press, related to prominent applications :

- automated decisions for autonomous vehicles
- loan agreements
- Admission Post Bac/ParcourSup

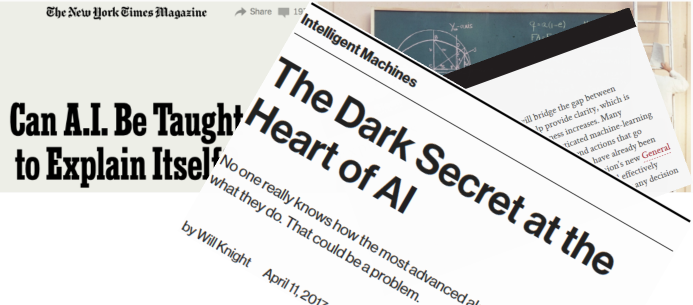


# Motivations

- new regulations (eg. GDPR)
- raising concern in the society : making A.I. systems trustable !

Featured in mainstream press, related to prominent applications :

- automated decisions for autonomous vehicles
- loan agreements
- Admission Post Bac/ParcourSup



# General Data Protection Regulation : A right to explanation ?

However, in their examination of the legal status of the GDPR, Wachter et al. conclude that such a right does not exist yet. The right to explanation is only explicitly stated in a recital :

> *a person who has been subject to automated decision-making "should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision "*

However, recitals are not legally binding. It also appears to have been intentionally not included in the final text of the GDPR after appearing in an earlier draft.

# General Data Protection Regulation : A right to explanation ?

Still, Article 13 and 14 about notification duties may provide a right to be informed about the “logic involved” prior to decision

> *“existence of automated decision-making, including profiling [...] [and provide data subjects with] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing.”*

As it stands, only provides a (limited : secret of affairs, etc.) right to obtain ex-ante explanations about the model (which they call, ‘right to be informed’).

Wachter et al. *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.* International Data Privacy Law, 2017.

# Loi pour une république numérique

L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

- Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
- Les données traitées et leurs sources ;
- Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
- Les opérations effectuées par le traitement.

Décret du 14 Mars 2017, cité et commenté dans :

Besse et al.. *Loyauté des Décisions Algorithmiques.* Contribution to CNIL debate, 2017.

# Transparency, Interpretability or Explainability ?

According to Besse et al., a decision can be said to be :

- transparent when the algorithm/code are made available.
- interpretable when it is possible to identify the features or variables which were prominent for the decision (even sometimes quantify this importance)
- explainable when it is possible to explicitly relate the values taken by the input data and the taken decision

Besse et al.. *Loyauté des Décisions Algorithmiques*. Contribution to CNIL debate, 2017 (my translation).

# Transparency does not imply explainability

```
(lambda _, __, ___, ____, _____, ______, _______, ________:
    getattr(
        __import__(True.__class__.__name__[_] + [].__class__.__name__[__]),
        ().__class__.__eq__.__class__.__name__[:__] +
        ().__iter__().__class__.__name__[_____:________]
    )(
        _, (lambda _, __, ___: _(_, __, ___))(
            lambda _, __, ___:
                chr(___ % __) + _(_, __, ___ // __) if ___ else
                (lambda: _).func_code.co_lnotab,
            _ << ________,
            (((_____ << ____) + _) << ((___ << _____) - ___)) + (((((___ << __)
            - _) << ___) + _) << ((_____ << ____) + (_ << _))) + (((_______ <<
            __) - _) << (((((_ << ___) + _)) << ___) + (_ << _))) + (((_______
            << ___) + _) << ((_ << ______) + _)) + (((_______ << ____) - _) <<
            ((_______ << ___))) + (((_ << ____) - _) << ((((___ << __) + _) <<
            __) - _)) - (_______ << ((((___ << __) - _) << __) + _)) + (_______
            << (((((_ << ___) + _)) << __))) - ((((((_ << ___) + _)) << __) +
            _) << ((((___ << __) + _) << _))) + (((_______ << __) - _) <<
            (((((_ << ___) + _)) << _))) + (((___ << ___) + _) << ((_____ <<
            _))) + (_____ << ______) + (_ << ___)
        )
    )
)(
    *(lambda _, __, ___: _(_, __, ___))(
        (lambda _, __, ___:
            [__(___[(lambda: _).func_code.co_nlocals])] +
            _(_, __, ___[(lambda _: _).func_code.co_nlocals:]) if ___ else []
        ),
        lambda _: _.func_code.co_argcount,
        (
            lambda _: _,
            lambda _, __: _,
            lambda _, __, ___: _,
            lambda _, __, ___, ____: _,
            lambda _, __, ___, ____, _____: _,
            lambda _, __, ___, ____, _____, ______: _,
            lambda _, __, ___, ____, _____, ______, _______: _,
            lambda _, __, ___, ____, _____, ______, _______, ________: _
        )
    )
)
```

# Transparency does not imply explainability

```python
(lambda _, __, ___, ____, _____, ______, _______, ________:
    getattr(
        __import__(True.__class__.__name__[_] + [].__class__.__name__[__]),
        ().__class__.__eq__.__class__.__name__[:__] +
        ().__iter__().__class__.__name__[_____:________]
    )(
        _, (lambda _, __, ___: _(_, __, ___))(
            lambda _, __, ___:
                chr(___ % __) + _(_, __, ___ // __) if ___ else
                (lambda: _).func_code.co_lnotab,
            _ << ________,
            (((_____ << ____) + _) << ((___ << _____) - ___)) + (((((___ << __)
            - _) << ___) + _) << ((_____ << ____) + (_ << _))) + (((_______ <<
            __) - _) << (((((_ << ___) + _)) << ___) + (_ << _))) + (((_______
            << ___) + _) << ((_ << ______) + _)) + (((_______ << ____) - _) <<
            ((_______ << ___))) + (((_ << ____) - _) << ((((___ << __) + _) <<
            __) - _)) - (_______ << ((((___ << __) - _) << __) + _)) + (_______
            << (((((_ << ___) + _)) << __))) - ((((((_ << ___) + _)) << __) +
            _) << ((((___ << __) + _) << _))) + (((_______ << __) - _) <<
            (((((_ << ___) + _)) << _))) + (((___ << ___) + _) << ((_____ <<
            _))) + (_____ << ______) + (_ << ___)
        )
    )
)(
    *(lambda _, __, ___: _(_, __, ___))(
        (lambda _, __, ___:
            [__(___[(lambda: _).func_code.co_nlocals])] +
            _(_, __, ___[(lambda _: _).func_code.co_nlocals:]) if ___ else []
        ),
        lambda _: _.func_code.co_argcount,
        (
            lambda _: _,
            lambda _, __: _,
            lambda _, __, ___: _,
            lambda _, __, ___, ____: _,
            lambda _, __, ___, ____, _____: _,
            lambda _, __, ___, ____, _____, ______: _,
            lambda _, __, ___, ____, _____, ______, _______: _,
            lambda _, __, ___, ____, _____, ______, _______, ________: _
        )
    )
)
```

prints `Hello World!` (by Ben Kurtovic, winner of a 2017 *obfuscation contest)*

# A panel of questions we need to answer?

1. what were the main factors in a decision?
2. would changing a given factor have changed the decision?
3. how to improve the decision?
4. why did two similar-looking cases get different conclusions, or vice-versa?
5. does the model indeed do what is expected?
6. why this decision (recommendation)?
7. ...

# The explanation landscape is rich already

## Examples of approaches

- data-based explanations (incl. counterfactuals) [Datta et al., 2016]
- locally faithful approximations (LIME), surrogate models [Ribeiro et al, 2016]
- add constraints or objective (capturing interpretability) [Sokolovska et al., 2017] ;
- restrict operators to argumentation schemes validated by the user. [Belahcène et al., 2017]
- ...

Datta et al.. *Algorithmic transparency via quantitative input influence : Theory and experiments with learning systems.* The 37th IEEE Symposium on Security and Privacy.2016.

Ribeiro et al.. *"why should i trust you ?" Explaining the predictions of any classifier.* In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.2016.

Sokolovska et al.. *The fused lasso penalty for learning interpretable medical scoring systems.2017.* IJCNN.

Belahcène et al.. *Explaining robust additive utility models by sequences of preference swaps.* Theory and Decision. 2017.

# The explanation landscape is rich already

## Examples of approaches

- data-based explanations (incl. counterfactuals) [Datta et al., 2016]
- locally faithful approximations (LIME), surrogate models [Ribeiro et al, 2016]
- add constraints or objective (capturing interpretability) [Sokolovska et al., 2017] ;
- restrict operators to argumentation schemes validated by the user. [Belahcène et al., 2017]
- ...

## An explanation (argumentation) scheme

an operator tying a *tuple of premises* (pieces of information provided or approved by the Decision Maker, or inferred during the process, and some supplementary hypotheses on the reasoning process (model's assumptions) to *a conclusion.*

# Contents

Motivations

Introduction to Multiple Criteria Decision Aiding
- Basic MCDA concepts
- Preference Elicitation

Explanation schemes in MCDA context
- Pairwise comparisons
- Ordinal Sorting

Future prospects and applications

# Our context : Multiple Criteria Decision Aiding

A **performance table**, describing several actions according to various criteria - the higher the better

A **decision problem** : Is action A better than action B ? Is action C good enough ?

Sparse **preferences** between some actions

**Decision Maker**

# Pairwise comparisons (choice or ranking)

I want to compare hotels described by 4 criteria :

| | | | | |
|---|---|---|---|---|
| $\mathcal{A}$ - | comfort : | ($4^\star$) | A $\succ$ a | ($2^\star$) |
| $\mathcal{B}$ - | restaurant : | (presence) | B $\succ$ b | (absence) |
| $\mathcal{C}$ - | commute time : | (15 min) | C $\succ$ c | (45 min) |
| $\mathcal{D}$ - | cost : | (50 $) | D $\succ$ d | (150 $) |

I **prefer** [AbCd] to [aBcD], [abcD] to [aBCd] and [aBCd] to [Abcd]

I want to know :
Is [abCD] better than [ABcd] ?

**Decision Maker**

# ORDINAL SORTING

| Object | a | b | c | d | Assignment |
|---|---|---|---|---|---|
| $A_1$ | 3 | 3 | 2.5 | 0 | ⋆⋆⋆ |
| $A_2$ | 3 | 2 | 2.1 | 1 | ⋆⋆⋆ |
| $B_1$ | 2 | 2 | 1.3 | 1 | ⋆⋆ |
| $B_2$ | 3 | 1 | 3.7 | 0 | ⋆⋆ |
| $C_1$ | 2 | 1 | 1.6 | 1 | ⋆ |
| $C_2$ | 1 | 1 | 4.1 | 0 | ⋆ |
| X | 2 | 2 | 1.1 | 0 | ? |

**What class should I assign to X?**

**Decision Maker**

# Our context : Multiple Criteria Decision Aiding

**Decision Maker**

A **performance table**, describing several actions according to various criteria - the higher the better

A **decision problem** : Is action A better than action B ? Is action C good enough ?

Sparse **preferences** between some actions

A **recommendation**

ANALYST

# Our context : Multiple Criteria Decision Aiding


A **performance table**, describing several actions according to various criteria - the higher the better

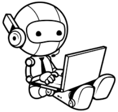
A **decision problem** : Is action A better than action B? Is action C good enough?

Sparse **preferences** between some actions

A **recommendation**

Assumes a **preference model** containing **aggregation procedures**

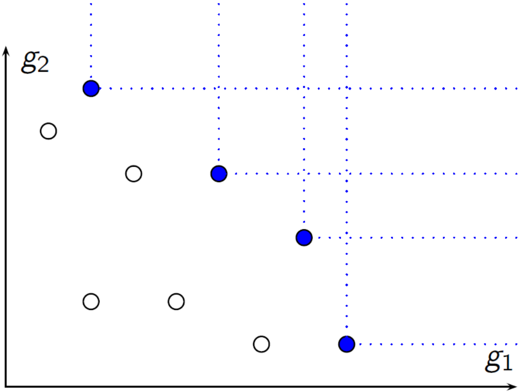
- mapping feature profiles to recommendations.
- extending Pareto dominance and expressed preferences.
- implementing a decision theoretic stance.



# Dominance, Pareto-optimality

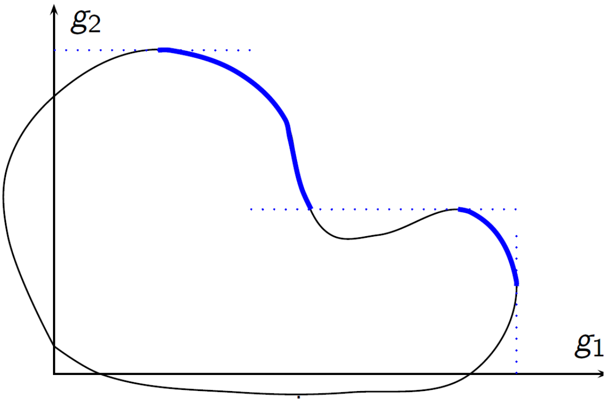
- Consider $a = (a_1, a_2, \ldots, a_n)$, $b = (b_1, b_2, \ldots, b_n)$,
  $a \Delta b$ iff $a_j \geq b_j, \forall j = 1..n$, one of the inequalities being strict,
- The dominance relation $\Delta$ expresses unanimity among criteria in favor of one action in the comparison,
- $\Delta$ defines on $A$ strict partial order (asymmetric and transitive),
- $\Delta$ is usually very poor,
- $a \in A$ is Pareto-optimal iff $\nexists b \in A$ s.t. $b \Delta a$,

# Pareto front




Pareto front in a discret bi-criteria problem

# Pareto front




Pareto front in a continuous bi-critera problem

# Preference information

- To discriminate among Pareto-optimal alternatives, the dominance relation $\Delta$ is useless,
- Decision aiding requires to enrich $\Delta$ by additional information called **preference information**,
- Preference information refers to the DM's opinions, value system, convictions ... concerning the decision problem,
- It is standard to distinguish :
  - Intracriterion preference information, and
  - Intercriteria preference information.

# MCDA

## Model selection

- a preference model contains aggregation procedures satisfying common properties.
- a model is selected considering decision stance, expressiveness, tractability.

### Additive Utility Model

- preference derives from a value model

$$\exists V \; s.t. \; x \succsim y \iff V(x) \geq V(y)$$

- value is additive (i.e. $V(x) = \sum_i v_i(x_i)$)

### NonCompensatory Sorting Model

- pairwise comparisons preferences

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

# MCDA

## Model elicitation

- Once a model is selected, a specific decision procedure has to be determined.
- preference information is collected from the Decision Maker, then processed.

| Approach | Summary | Pros | Cons |
|---|---|---|---|
| Complete | Standard sequence of questions | Unequivocal | Demanding |
| Partial | Learning from DM's statements + Loss function | Efficient | Arbitrary |
| Robust | Partial + Accounting for possible completions | Cautious | Indecisive |
| Active | Dynamically determined queries minimizing regret | Fast | Arbitrary |

# Contents

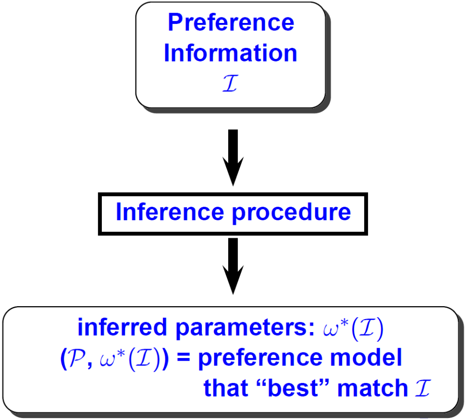
Motivations

Introduction to Multiple Criteria Decision Aiding
- Basic MCDA concepts
- Preference Elicitation

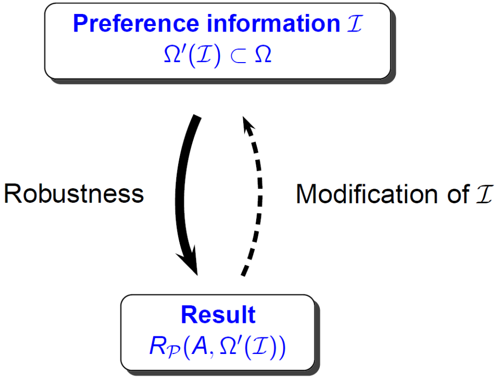
Explanation schemes in MCDA context
- Pairwise comparisons
- Ordinal Sorting

Future prospects and applications

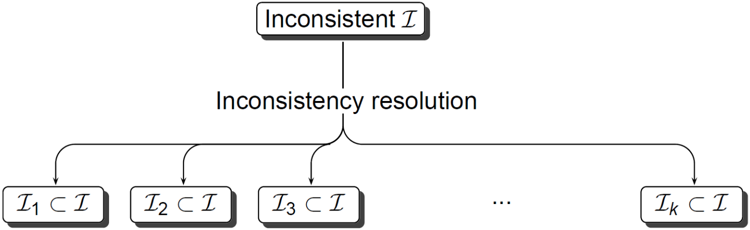
# Preference Elicitation



# PREFERENCE ELICITATION



# Preference Elicitation



# Our context : Multiple Criteria Decision Aiding

A **performance table**, describing several actions according to various criteria - the higher the better

A **decision problem** : Is action A better than action B ? Is action C good enough ?

Sparse **preferences** between some actions

A **recommendation**

Why ?

A Recommendation Supported with an **explanation** :

Expressed Preference + Model Features $\Rightarrow$ Recommendation

**Decision Maker**

ANALYST

# Contents

Motivations

Introduction to Multiple Criteria Decision Aiding
- Basic MCDA concepts
- Preference Elicitation

Explanation schemes in MCDA context
- Pairwise comparisons
- Ordinal Sorting

Future prospects and applications

# ?

I want to compare hotels described by 4 criteria : comfort (A), parking (B), commute time (C), and cost (D).

I **prefer** :
( 4⋆, no, 15 min, 150 \$) to ( 2⋆, yes, 45 min, 50 \$),
( 2⋆, no, 45 min, 50 \$) to ( 2⋆, yes, 15 min, 150 \$),
( 2⋆, yes, 15 min, 150 \$) to ( 4⋆, no, 45 min, 150 \$).

I want to know :
Is (2⋆, no, 15 min, 50 \$) better than (4⋆, yes, 45 min, 150 \$)?

**Decision Maker**

ANALYST

Assumptions :

- preference derives from a value model (i.e. $\exists V$ s.t. $x \succsim y \iff V(x) \geq V(y)$)
- value is additive (i.e. $V(x) = \sum_i v_i(x_i)$)

# ?

Decision Maker

I want to compare hotels described by 4 criteria : comfort (A), parking (B), commute time (C), and cost (D).

I **prefer** :
( 4⋆, no, 15 min, 150 \$) to ( 2⋆, yes, 45 min, 50 \$),
( 2⋆, no, 45 min, 50 \$) to ( 2⋆, yes, 15 min, 150 \$),
( 2⋆, yes, 15 min, 150 \$) to ( 4⋆, no, 45 min, 150 \$).

I want to know :
Is (2⋆, no, 15 min, 50 \$) better than (4⋆, yes, 45 min, 150 \$) ?

ANALYST

Ordinal encoding : attribute values of interest are sorted and encoded

criterion A : 4⋆ is strong (⦿), 2⋆ is weak (○) ; criterion B : yes is strong (⦿), no is weak (○) ; criterion C : 15 min is strong (⦿), 45 min is weak (○) ; criterion D : 50 \$ is strong (⦿), 150 \$ is weak (○).

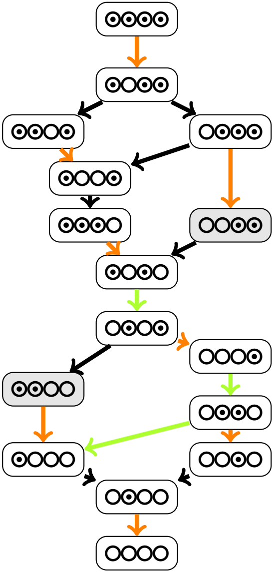
# Streamlining the Robust Additive Value model

A B C D / A B C D

I **prefer** [⦿○⦿○] to [○⦿○⦿], [○○○⦿] to [○⦿⦿○] and [○⦿⦿○] to [⦿○○○]

I want to know : Is [○○⦿⦿] better than [⦿⦿○○] ?

It is !



## Knowledge representation

- comparative pairwise statements are represented as inequations between elementary score differences
- Knowledge Base : Preference Information (↘) + Pareto dominance (↘)

## Inference

- a comparative statement holds iff it is a conical combination of statements in the KB
- a finite number of inferred statements (↘) are computed by Linear Programming

# Explaining with sequences of preference swaps

- Assuming the complexity of preference stems from having many moving parts
- Decomposing the complexity into smaller grains by reasoning *ceteris paribus*

☞ explanations can be long, but can be kept short and computed efficiently when constraining the PI

**I prefer** [⦿○⦿○] to [○⦿○⦿], [○○○⦿] to [○⦿⦿○] and [○⦿⦿○] to [⦿○○○]
**I want to know :** is [○○⦿⦿] better than [⦿⦿○○] ?

Decision Maker

It is ! Here is why :

1. [○, ○, ⦿, ⦿] is better than [⦿, ○, ⦿, ○] because, everything else being equal, [○, b, c, ⦿] (2★ for 50 $) is better than [⦿, b, c, ○] (4★ for 150 $).
2. [⦿, ○, ⦿, ○] is better than [○, ⦿, ○, ⦿] because, you told me so!
3. [⦿, ○, ⦿, ○] is better than [⦿, ⦿, ○, ○] because, everything else being equal, [a, ○, ⦿, d] (no parking, 15 min commute) is better than [a, ⦿, ○, d] (parking, 45 min)

analyst

Belahcène et al. *Explaining robust additive utility models by sequences of preference swaps.* Theory and Decision. 2017.

# Contents

Motivations

Introduction to Multiple Criteria Decision Aiding
- Basic MCDA concepts
- Preference Elicitation

Explanation schemes in MCDA context
- Pairwise comparisons
- Ordinal Sorting

Future prospects and applications

# NonCompensatory Sorting procedure

## Output

- A category among an ordered set $C_1 \prec \cdots \prec C_p$

## Sorting rule

- an alternative is in category $C_h$ or better iff it has sufficient attributes at level $C_h$ on a coalition of criteria deemed sufficient at level $C_h$

## History

- inspired by Electre Tri
- described and characterized in [Bouyssou & Marchant, 2007 ab]
- equivalent to the Sugeno integral model [Slowinski et al., 2002]

## Particular cases

- **U** : using a Unique set of sufficient coalitions of criteria
- **V** : representing sufficient coalitions with a Voting model
- We call NCS models following **U** "U-NCS", **U&V** "MR-Sort"[Leroy et al., 2011]

# NonCompensatory Sorting Example

| Project | a | b | c | d | Category |
|---|---|---|---|---|---|
| $p_1$ | 5 | 6 | 6 | 5 | ? |
| $p_2$ | 3.5 | 1 | 3 | 9 | ? |
| $p_3$ | 7.5 | 2 | 1 | 3 | ? |
| $p_4$ | 2 | 8 | 2.5 | 7 | ? |
| $p_5$ | 3 | 8.5 | 3 | 8.5 | ? |
| $p_6$ | 8 | 4 | 1.5 | 1.5 | ? |

| | | | | | |
|---|---|---|---|---|---|
| $\star$ | $< 4$ | $< 3$ | $< 2$ | $< 2$ | boundary between $\star$ and $\star\star$ |
| $\star\star$ | $[4,7[$ | $[3,8[$ | $[2,5[$ | $[2,8[$ | |
| $\star\star\star$ | $\geq 7$ | $\geq 8$ | $\geq 5$ | $\geq 8$ | boundary between $\star\star$ and $\star\star\star$ |

# NonCompensatory Sorting Example

## $1^{st}$ phase : criterion-wise sorting

| project | a | b | c | d | Category |
|---|---|---|---|---|---|
| $p_1$ | $\star\star$ | $\star\star$ | $\star\star\star$ | $\star\star$ | ? |
| $p_2$ | $\star$ | $\star$ | $\star\star$ | $\star\star\star$ | ? |
| $p_3$ | $\star\star\star$ | $\star$ | $\star$ | $\star\star$ | ? |
| $p_4$ | $\star$ | $\star\star\star$ | $\star\star$ | $\star\star$ | ? |
| $p_5$ | $\star$ | $\star\star\star$ | $\star\star$ | $\star\star\star$ | ? |
| $p_6$ | $\star\star\star$ | $\star\star$ | $\star$ | $\star$ | ? |

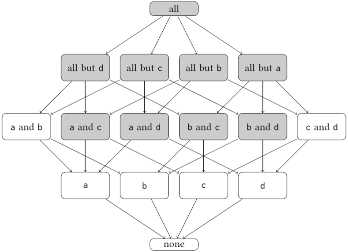
| | | | | | |
|---|---|---|---|---|---|
| $\star$ | $< 4$ | $< 3$ | $< 2$ | $< 2$ | boundary between $\star$ and $\star\star$ |
| $\star\star$ | [4,7[ | [3,8[ | [2,5[ | [2,8[ | |
| $\star\star\star$ | $\geq 7$ | $\geq 8$ | $\geq 5$ | $\geq 8$ | boundary between $\star\star$ and $\star\star\star$ |

# NonCompensatory Sorting Example

## $2^{nd}$ phase : noncompensatory multi criteria aggregation

| project | a | b | c | d | Category |
|---|---|---|---|---|---|
| $p_1$ | ⋆⋆ | ⋆⋆ | ⋆ ⋆ ⋆ | ⋆⋆ | ? |
| $p_2$ | ⋆ | ⋆ | ⋆⋆ | ⋆ ⋆ ⋆ | ? |
| $p_3$ | ⋆ ⋆ ⋆ | ⋆ | ⋆ | ⋆⋆ | ? |
| $p_4$ | ⋆ | ⋆ ⋆ ⋆ | ⋆⋆ | ⋆⋆ | ? |
| $p_5$ | ⋆ | ⋆ ⋆ ⋆ | ⋆⋆ | ⋆ ⋆ ⋆ | ? |
| $p_6$ | ⋆ ⋆ ⋆ | ⋆⋆ | ⋆ | ⋆ | ? |

**Sufficient coalitions**




**Insufficient coalitions**

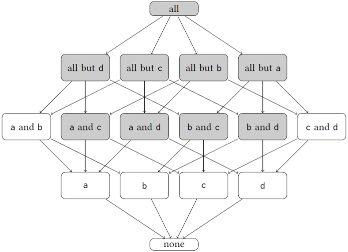
- Getting an overall ⋆⋆ or ⋆ ⋆ ⋆ requires getting ⋆⋆ or ⋆ ⋆ ⋆ on a sufficient coalition of criteria
- Getting an overall ⋆ ⋆ ⋆ requires getting ⋆ ⋆ ⋆ on a sufficient coalition of criteria

# NonCompensatory Sorting Example

## $2^{nd}$ phase : noncompensatory multi criteria aggregation

| project | a | b | c | d | Category |
|---|---|---|---|---|---|
| $p_1$ | ★★ | ★★ | ★★★ | ★★ | ★★ |
| $p_2$ | ★ | ★ | ★★ | ★★★ | ★ |
| $p_3$ | ★★★ | ★ | ★ | ★★ | ★★ |
| $p_4$ | ★ | ★★★ | ★★ | ★★ | ★★ |
| $p_5$ | ★ | ★★★ | ★★ | ★★★ | ★★★ |
| $p_6$ | ★★★ | ★★ | ★ | ★ | ★ |

**Sufficient coalitions**




**Insufficient coalitions**

- Getting an overall ★★ or ★★★ requires getting ★★ or ★★★ on a sufficient coalition of criteria
- Getting an overall ★★★ requires getting ★★★ on a sufficient coalition of criteria

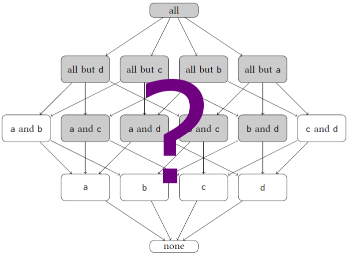
# Learning / Disaggregation of U-NCS model

☞ **Input : profiles + reference assignments**

| model | a | b | c | d | Category |
|---|---|---|---|---|---|
| $m_1$ | 16 973 | 29 | 2.66 | 2.5 | ⋆⋆ |
| $m_2$ | 18 342 | 30.7 | 2.33 | 3 | ⋆ |
| $m_3$ | 15 335 | 30.2 | 2 | 2.5 | ⋆⋆ |
| $m_4$ | 18 971 | 28 | 2.33 | 2 | ⋆⋆ |
| $m_5$ | 17 537 | 28.3 | 2.33 | 2.75 | ⋆ ⋆ ⋆ |
| $m_6$ | 15 131 | 29.7 | 1.66 | 1.75 | ⋆ |

| | | | | | |
|---|---|---|---|---|---|
| ⋆/ ⋆⋆ | ? | ? | ? | ? | |
| ⋆⋆/ ⋆ ⋆ ⋆ | ? | ? | ? | ? | |

**Sufficient coalitions**




**Insufficient coalitions**

☞ **Expected Outputs : set of profiles + set of sufficient coalitions.**

# Learning / Disaggregation of U-NCS model

- Direct elicitation with standard sequence procedures
- Computational issues with the indirect elicitation of MR Sort (learning from assignment examples) :
  - with a MIP [Leroy et al, 2011] : hardly more than toy examples
  - with a meta-heuristic [Sobrie et al, 2016] : learning sets from preference learning
- issues with knowledge representation
  - dependencies between profiles and coalitions are non-trivial
  - the profiles part seems to fall within the domain of 'logical inference'
  - the coalition part is described by linear programming

+ need for a unified description : back to NCS (alternate solution : MR Sort + cutting planes ?)

# A compact SAT formulation

Let $\alpha : \mathbb{X} \to \{\textsc{Good}, \textsc{Bad}\}$ an assignment. We define the boolean function $\phi_\alpha^{\mathit{pairwise}}$ 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 $g \in \alpha^{-1}(\textsc{Good})$ and a *bad* alternative $b \in \alpha^{-1}(\textsc{Bad})$,

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

$$\begin{array}{lcc}
\phi_\alpha^1 := & \bigwedge_{i \in \mathcal{N}} \bigwedge_{x' \succsim_i x} & (\lambda_{i,x'} \vee \neg \lambda_{i,x}) \\
\phi_\alpha^2 := & \bigwedge_{i \in \mathcal{N},\ g \in \alpha^{-1}(\textsc{Good}),\ b \in \alpha^{-1}(\textsc{Bad})} & (\neg \mu_{i,g,b} \vee \neg \lambda_{i,b}) \\
\phi_\alpha^3 := & \bigwedge_{i \in \mathcal{N},\ g \in \alpha^{-1}(\textsc{Good}),\ b \in \alpha^{-1}(\textsc{Bad})} & (\neg \mu_{i,g,b} \vee \lambda_{i,g}) \\
\phi_\alpha^4 := & \bigwedge_{g \in \alpha^{-1}(\textsc{Good}),\ b \in \alpha^{-1}(\textsc{Bad})} & (\bigvee_{i \in \mathcal{N}} \mu_{i,g,b})
\end{array}$$

# Towards explanations for NCS

## Situation 1 : Auditing conformity

An independent audit agency is commissioned to check that the decision on the the committee indeed comply with a publicly announced decision rule.

# Towards explanations for NCS

## Situation 1 : Auditing conformity

An independent audit agency is commissioned to check that the decision on the the committee indeed comply with a publicly announced decision rule.

☞ computing and providing a certificate of feasibility of a SAT problem.

# Towards explanations for NCS

## Situation 1 : Auditing conformity

An independent audit agency is commissioned to check that the decision on the the committee indeed comply with a publicly announced decision rule.

☞ computing and providing a certificate of feasibility of a SAT problem.

## Situation 2 : Justifying individual decisions

A candidate, (supposedly) unsatisfied with the outcome of the process regarding his own classification, challenged the committee and asks for justification.

- necessary decisions entailed by the jurisprudence.
- Ambivalent situations.

# Towards explanations for NCS

## Situation 1 : Auditing conformity

An independent audit agency is commissioned to check that the decision on the the committee indeed comply with a publicly announced decision rule.

☞ computing and providing a certificate of feasibility of a SAT problem.

## Situation 2 : Justifying individual decisions

A candidate, (supposedly) unsatisfied with the outcome of the process regarding his own classification, challenged the committee and asks for justification.

- necessary decisions entailed by the jurisprudence.
- Ambivalent situations.

☞ computing and providing a certificate of infeasibility (MUS)

# Towards explanations for NCS

**Open issues :**

- How do we leverage this description inside a decision process ?
- Can we build explanations around certificates of UNSAT (MUSes) ?
    - What is a "good" certificate ?
    - Can we find a template (=argument schemes) in which they fit ? (all of them ? some of them ?)
    - Can we compute them effectively ?

# FUTURE PROSPECTS AND APPLICATIONS

## Open issues

- Intégration de l'explication et de l'élicitation dans un mécanisme dialectique (gestion de l'inconsitance, choix de modèle, protocole de dialogue, etc. )
  - PEPS "PULP" (S. Destercke, Heudiasyc - Lip6)
  - Propale ANR 2018 IRELAND" (V. Mousseau / W. Ouerdane, LGI - LIP6 - LAMSADE- IMT Atlantique)
- Encodages et méthodes SAT pour la production d'explications.
  - PEPS "SAT4EX" (N. Maudet, Lip6 - CRIL)
- ...

## Different application domains

- Configuration problem ;
- Recommendation problem ;
- Administrative decisions ;
- ...