Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value

Christophe Labreuche, Simon Fossier
Thales Research & Technology, 1 avenue Fresnel, 91767 Palaiseau cedex, France
{christophe.labreuche,simon.fossier}@thalesgroup.com

Abstract

The capability to explain the result of aggregation models to decision makers is key to reinforcing user trust. In practice, Multi-Criteria Decision Aiding models are often organized in a hierarchical way, based on a tree of criteria. We present an explanation approach usable with any hierarchical multi-criteria model, based on an *influence index* of each attribute on the decision. A set of desirable axioms are defined. We show that there is a unique index fulfilling these axioms. This new index is an extension of the Shapley value on trees. An efficient rewriting of this index, drastically reducing the computation time, is obtained. Finally, the use of the new index is illustrated on an example.

1 Introduction

Decision Aiding plays a central role in many systems and processes, ranging from recommender systems to system engineering. Any complex decision process is characterized by the presence of multiple and often conflicting criteria. To facilitate this process, Multi-Criteria Decision Aiding (MCDA) proposes a wide variety of models to aggregate the criteria, and dedicated methods to respect the preferences of the decision maker (DM). The criteria are usually organized hierarchically with several nested aggregation functions, in order to represent the natural decomposition of the decision reasoning into points of view and sub-points of view. The next example presents a hierarchical MCDA model.

Example 1 *The mission of Maritime Patrol is to monitor a maritime area and seek for illegal activity. It evaluates in real time a Priority Level (PL) associated to each ship in this area, where the PL increases when there is a suspicion of illegal activity or when it is urgent to intercept the ship. The PL is intrinsically based on multiple criteria: 1. Incoherence between Automatic Identification System (AIS) data and radar detection; 2. Suspicion of drug smuggling on the ship; 3. Suspicion of human smuggling on the ship; 4. Current speed (since fast boats are often used to avoid being easily intercepted); 5. Maximum speed since the first detection of the ship (it represents the urgency for the potential interception); 6. Proximity of the ship to the shore (since smuggling ships often aim at reaching the shore as fast as possible).*

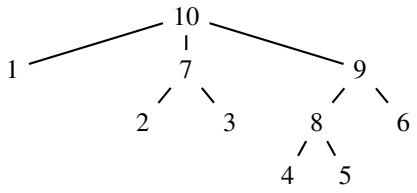

Figure 1: Hierarchy of criteria for Ex. 1.

In this running example, the criteria will only take intuitive values 'High/Yes', which correspond to higher PL, or 'Low/None/No'. More intermediate or continuous values can be defined without any loss of generality. According to the operational commander (OC), the six criteria are organized as in Fig. 1. There are four aggregation nodes, with the following semantics: 7. Suspicion of illegal activity; 8. Kinematics; 9. Capability to escape interception; 10. Overall PL. ■

The OC has the following needs:

N1 (Interpretability): In order to trust the PL model, the OC needs to have a general interpretation of the PL model: What are the most important attributes on average?

N2 (Explicability): Beyond the PL score itself, the OC needs further analysis to determine whether to intercept a ship: Why is the PL higher for this ship? Why has the PL of this ship significantly increased over last minutes?

N3 (Sensibility Analysis): For a ship with a low PL score, the OC wonders what changes in the attribute values would increase the PL most significantly. The maritime patrol will then put more means in identifying as early as possible whether these identified attributes vary.

The answers to the previous questions require *explanation* capabilities of the MCDA model. Explanations are also important to build acceptance and trust for all users.

The OC is not looking for complete explanations (see [Miller, 2017]) – such as proofs – as the OC is under stress and time pressure, and is not specialist in MCDA (and is unable to follow a complex MCDA reasoning). On the contrary, we are aiming at simple and incomplete explanations, such as: *the PL has significantly increased mainly because of 'AIS/radar incoherence' status change, and not because of the increasing current speed*. Such explanations can be derived from sensitivity analyses [Ribeiro *et al.*, 2016]. To this end, we construct an *index* measuring the *influence* of each

attribute on the decision (e.g. the comparison between the PL for two ships). In the previous example, the influence of attribute 1 would be important and that of attribute 4 would be small.

The influence index can be defined in many different ways. The existing proposals only consider a flat organization of criteria, and are defined for specific models (see Section 3). In order to derive the correct expression, we adopt an axiomatic approach, thereby defining a set of desirable properties (axioms) for such indices. Some of these axioms are related to the tree structure of the criteria. This new index is an extension of the Shapley value (defined in Cooperative Game Theory) on trees. Note that the use of values has recently gained interest for interpretability in Classification [Datta *et al.*, 2016; Lundberg and Lee, 2017].

The main contributions of this paper are now summarized:

- Our influence index is formally justified by an axiomatic approach, where we have shown that it is the only index that fulfills some natural properties (Sect. 4 and 5);
- Our influence index is easily interpretable as it can be derived from an intuitive construction (Sect. 5);
- Despite of exponential complexity, its computation can be drastically reduced by considering a reduced tree, making it usable in practical applications (Sect. 6);
- Its application is relevant on the running example for the three needs **N1**, **N2** and **N3**, and we illustrate its benefits over the existing proposals in an example (Sect. 7).

2 Model and Notations

2.1 MCDA Model

Multi-Criteria Decision Aiding (MCDA) models aim at representing the preferences of a Decision Maker (DM) over a set of criteria $N = \{1, \dots, n\}$, each criterion $i \in N$ being associated with an attribute X_i, either discrete or continuous. The alternatives are characterized by a value on each attribute and are thus associated to an element in $X = X_1 \times \cdots \times X_n$. Even if some elements of X may not correspond to real alternatives, we will call X the set of alternatives. We assume that the preferences of the DM over X are represented by a cardinal utility model $U : X \to \mathbb{R}$. It is classical to write U in the *decomposable* way [Keeney and Raiffa, 1976]:

$$U(x) = H(u_1(x_1), \dots, u_n(x_n)) \qquad \forall x \in X, \qquad (1)$$

where the $u_i : X_i \to \mathbb{R}$ are called the *utility functions* and $H : \mathbb{R}^n \to \mathbb{R}$ is an aggregation function.

2.2 Structuring Criteria in a Hierarchy

In many applications, the number of criteria can be relatively large – beyond 20 or more. It is then common to decompose H in a hierarchical way, with intermediate aggregation nodes. This structure helps to keep the number of aggregated elements to a reasonable number for each aggregation node (see the Miller law [Miller, 1956]), so that the DM can easily express preferences to elicit each aggregation function separately, as the OC would do in Ex. 1.

The hierarchy of criteria is represented by a rooted tree T, defined by the set of nodes M_T (i.e. the set of criteria and aggregation nodes), and the children $\mathrm{Ch}_T(l)$ of node l (i.e. the nodes that are aggregated at each node l) [Diestel, 2005]. We also denote by $N_T \subseteq M_T$ the set of leaves of tree T (i.e. the criteria), by $s_T \in M_T$ the root of tree T (i.e. the top aggregation node), by $\mathrm{Desc}_T(l)$ the set of descendants of l, and by $\mathrm{Leaf}_T(l)$ the leaves at or below $l \in M_T$. Unless specified, we assume that the leaves of tree T are exactly the criteria: $N_T = N$. We will consider two particular cases of trees: *flat organizations* of criteria, having only one aggregation node s_T (i.e. $M_T = N \cup \{s_T\}$), and *coalition structures* where each leaf is at depth exactly 2 to the root.

Representation (1) can be further decomposed on the tree structure T. For $x \in X$, we can compute $U(x)$ recursively from a function v_i^U defined at each node $i \in M_T$:

- $v_i^U(x) = u_i(x_i)$ for every leaf $i \in N_T$,
- $v_l^U(x) = H_l\left((v_k^U(x))_{k \in \mathrm{Ch}_T(l)}\right)$ for every aggregation node $l \in M_T \setminus N_T$, where H_l is an aggregation function,
- $U(x) = v_{s_T}^U(x)$ is the overall utility.

By abuse of notation, U will denote the vector $U := ((u_i)_{i \in N_T}, (H_i)_{i \in M_T \setminus N_T})$ of local functions characterizing U, as well as function yielding $U(x)$. We denote by $\mathcal{U}_T$ the set of such vectors U of local functions defined on T.

Example 2 (Ex. 1 continued) *We have* $M_T = \{1, \dots, 10\}$, $s_T = 10$ *and* $N_T = N = \{1, 2, 3, 4, 5, 6\}$. *The tree is defined by* $\mathrm{Ch}_T(7) = \{2, 3\}$ *(node* 7 *aggregates nodes* 2 *and* 3*),* $\mathrm{Ch}_T(8) = \{4, 5\}$, $\mathrm{Ch}_T(9) = \{6, 8\}$ *and* $\mathrm{Ch}_T(10) = \{1, 7, 9\}$. *Then we have* $U = (u_1, \dots, u_6, H_7, \dots, H_{10})$, *and* $U(x) = v_{10}^U(x)$, *with* $v_i^U(x) = u_i(x_i)$ *for all* $i \in N_T$, $v_7^U(x) = H_7(v_2^U(x), v_3^U(x))$, $v_8^U(x) = H_8(v_4^U(x), v_5^U(x))$, $v_9^U(x) = H_9(v_6^U(x), v_8^U(x))$, *and* $v_{10}^U(x) = H_{10}(v_1^U(x), v_7^U(x), v_9^U(x))$.

2.3 Examples of Aggregation Models

The goal of this paper is to design a generic explanation method, which can be applied to any model U and aggregation functions H_l. However, for illustrative purpose, we describe in this section two such models, classically used. Consider an aggregation node $l \in M_T \setminus N$, which children are $\mathrm{Ch}_T(l)$. For the sake of simplicity, we assume that the components that are aggregated by H_l are simply denoted by the vector $a = (a_1, \dots, a_{n_l})$, with $n_l = |\mathrm{Ch}_T(l)|$.

Weighted Sum. The weighted sum is a widely used aggregation function. It takes the following form

$$\mathrm{WS}(a) = \sum_{i=1}^{n_l} w_i \, a_i,$$

where w_i is the weight assigned to node (criterion) i. This model assumes the independence among the criteria. This assumption is often violated (presence of interaction among criteria), as we will see in the running example.

Choquet Integral. The Choquet integral [Choquet, 1953], which is an extension of the weighted sum, is a powerful aggregation function thanks to its ability to represent many types of interaction in a versatile way [Grabisch, 1996]. Interaction is materialized by non-linear terms. A very good compromise between the representation power of the model and

its elicitation and interpretation burden is obtained by restricting to interactions between pairs of criteria [Grabisch, 1997]. This is the 2-*additive* Choquet integral [Grabisch, 1997]:

$$\mathrm{CI}(a) = \sum_{i=1}^{n_l} w_i a_i + \sum_{1\le i<j\le n_l} \left(w_{ij}^{\wedge}\, a_i \wedge a_j + w_{ij}^{\vee}\, a_i \vee a_j \right) \quad (2)$$

where all coefficients $w_i, w_{ij}^{\wedge}, w_{ij}^{\vee}$ are non-negative and sum-up to one. The choice of min and max operators for the non-linear part (rather than other operators such as the product) comes from the fact that we need to satisfy idempotency (the overall score of an alternative having the same score $\alpha \in [0,1]$ on all criteria shall be equal to α, i.e. $\mathrm{CI}(\alpha,\dots,\alpha)=\alpha$). We note that the min and max operators are the numerical counterparts of the AND and OR connectives respectively.

Term $w_{ij}^{\wedge}$ (resp. $w_{ij}^{\vee}$) corresponds to *complementarity* (resp. *redundancy*) between criteria i and j in the sense that the better the achievement on one criterion, the more (resp. less) significant it is to improve on the other. This model is rich enough to capture most real life decision strategies.

Thanks to relation $a_i \wedge a_j + a_i \vee a_j = a_i + a_j$, (2) can be rewritten in a more compact form [Grabisch, 1997]:

$$\mathrm{CI}(a) = \sum_{i=1}^{n_l} m_i\, a_i + \sum_{1\le i<j\le n_l} m_{i,j}\, a_i \wedge a_j, \quad (3)$$

where $m_i = w_i + \sum_{j\neq i} w_{ij}^{\vee}$ and $m_{i,j} = w_{ij}^{\wedge} - w_{ij}^{\vee}$ are called the Möbius coefficients.

In order to interpret more easily these coefficients (e.g. to identify which criteria are the most important), the *mean importance* of criterion i is defined by [Grabisch, 1997]

$$\mathrm{Imp}_i(\mathrm{Ch}_T(l), m) = m_i + \sum_{j\neq i} \frac{m_{i,j}}{2}. \quad (4)$$

If m corresponds to a weighted sum (i.e. $m_{i,j}=0$ for all i,j), then $\mathrm{Imp}_i(\mathrm{Ch}_T(l),m)$ is the weight of criterion i. One can readily see that the Choquet integral is piecewise linear. More precisely, it is a weighted sum in each domain $\Sigma_\sigma = \{a\in[0,1]^{n_l} : a_{\sigma(1)} \le a_{\sigma(2)} \le \dots \le a_{\sigma(n_l)}\}$, where σ is a permutation on $\{1,\dots,n_l\}$:

$$\mathrm{CI}(a) = \sum_{i=1}^{n_l} a_i\, \Delta_i^a, \quad (5)$$

where $\Delta_i^a = m_i + \sum_{j\neq i,\, a_j>a_i} m_{i,j} + \sum_{j\neq i,\, a_j=a_i} \frac{m_{i,j}}{2}$. The previous three expressions (2), (3) and (5) will be used throughout this paper.

The next example illustrates this model.

Example 3 (Ex. 2 cont.) *After eliciting the OC's preferences, the aggregation functions are given by:*
Node 7**:** *There is suspicion of illegal activity whenever either drug or human smuggling is detected. Hence there is redundancy (operator* $\vee$*) between criteria* 2 *and* 3*. As human smuggling (crit.* 3*) is slightly more important than criterion* 2*, we obtain* $v_7^U(x) = 0.2\, v_3^U(x) + 0.8\, v_2^U(x) \vee v_3^U(x)$*;*
Node 8**:** $v_8^U(x) = (v_4^U(x) + v_5^U(x))/2$*;*
Node 9**:** *Nodes* 6 *and* 8 *are redundant (operator* $\vee$*), since there is a high risk that the ship escapes interception when it is either close to the shore (crit.* 6*) or very fast (node* 8*). Hence* $v_9^U(x) = 0.2\, v_6^U(x) + 0.2\, v_8^U(x) + 0.6\, v_6^U(x) \vee v_8^U(x)$*,*
Node 10**:** *Nodes* 1 *and* 7 *are redundant (operator* $\vee$*) since there is a suspicion on the ship when the score is high on either node* 1 *or* 7*. Nodes* 7 *and* 9 *are complementary (operator* $\wedge$*) as the risk is not so high for a suspicious ship (high value at node* 7*) that is easy to intercept (low value at node* 9*), or for a ship that is difficult to intercept but that is not suspicious. We have the same behavior between nodes* 1 *and* 9*. Hence* $v_{10}^U(x) = \big(v_1^U(x) \vee v_7^U(x) + v_1^U(x) \wedge v_9^U(x) + v_7^U(x) \wedge v_9^U(x)\big)/3$*.*

For x = (*Yes,None,None,High,High,High*)*, we obtain* $u_2(x) = u_3(x) = 0$*,* $u_i(x) = 1$ *for* $i \in \{1,4,5,6\}$*,* $v_7^U(x) = 0$*,* $v_8^U(x) = v_9^U(x) = 1$ *and* $U(x) = v_{10}^U(x) = \frac{2}{3}$*.*

Many other models can be used, such as the Generalized Additive Independence (GAI) model [Fishburn, 1967].

A tree can be reduced to a flat structure when all aggregation functions are weighted sums, thanks to linearity. This is no more true with the GAI model and the Choquet integral. Having a hierarchical structure is then a way to capture new aggregation functions not representable by a flat structure. For instance, the model of Ex. 3 cannot be rewritten as a single Choquet integral of the six attributes.

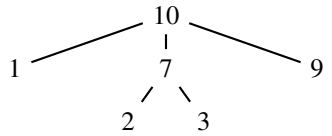

Figure 2: A subtree of Fig. 1. The details below node 9 are hidden.

2.4 Subtrees

In order to hide unnecessary details for some DMs, we consider subtrees of T having the same root. Let $\mathcal{T}_T$ be the set of trees T' such that $s_{T'} = s_T$, $M_{T'} \subseteq M_T$, and $\mathrm{Ch}_{T'}(l) = \mathrm{Ch}_T(l)$ for all $l \in M_{T'} \setminus N_{T'}$. The elements of $\mathcal{T}_T$ are interpreted as different levels of details on the original tree T. Fig. 2 illustrates this concept.

Given $U = ((u_i)_{i\in N_T}, (H_i)_{i\in M_T\setminus N_T}) \in \mathcal{U}_T$ and $T' \in \mathcal{T}_T$, we can define $U_{T'} = ((u'_i)_{i\in N_{T'}}, (H'_i)_{i\in M_{T'}\setminus N_{T'}}) \in \mathcal{U}_{T'}$ by $u'_i = u_i$ for $i \in N_{T'} \cap N_T$, $u'_i(x_i) = x_i$ for $i \in N_{T'} \setminus N_T$ and $H'_i = H_i$ for $i \in M_{T'} \setminus N_{T'}$. We set $X_i = \mathbb{R}$ for every $i \in M_T \setminus N_T$. Then for $x \in X$, $U(x) = U_{T'}(x^{T'})$ where $x^{T'} \in X_{T'}$ is defined by $x_i^{T'} = x_i$ if $i \in N_{T'} \cap N_T$ and $x_i^{T'} = v_i^U(x)$ otherwise.

Example 4 (Ex. 3 cont.) *For* T' *given by Fig. 2 and* x = (*Yes,None,None,High,High,High*)*, we obtain* $x_i^{T'} = x_i$ *for* $i \in \{1,2,3\}$*,* $x_9^{T'} = v_9^U(x) = 1$*,* $v_7^{U_{T'}}(x^{T'}) = 0$ *and* $U_{T'}(x^{T'}) = v_{10}^{U_{T'}}(x^{T'}) = \frac{2}{3}$*. Hence* $U_{T'}(x^{T'}) = U(x)$*.*

A particular subtree is when a node $j \in M_T$ of T becomes a leaf, and thus all descendants of j are encapsulated and represented by j. We define the restricted tree $T_{[j]}$ by $M_{T_{[j]}} := (M_T \setminus \mathrm{Desc}_T(j)) \cup \{j\}$, $N_{T_{[j]}} := (N_T \setminus \mathrm{Leaf}_T(j)) \cup \{j\}$, $s_{T_{[j]}} := s_T$, and $\mathrm{Ch}_{T_{[j]}}(l) = \mathrm{Ch}_T(l)$ for all $l \in M_{T_{[j]}} \setminus N_{T_{[j]}}$. Fig. 2 corresponds to $T_{[9]}$ where T is defined by Fig. 1.

For $J = \{j_1, \ldots, j_p\}$, we set $T_{[J]} := \Big(\big((T)_{[j_1]}\big)_{[j_2]} \cdots\Big)_{[j_p]}$.

For $x, y \in X_{N_{T'}}$ and $S \subseteq N_{T'}$, $(y_S, x_{-S}) \in X_{N_{T'}}$ denotes the alternative taking value y_i for $i \in S$ and value x_i for $i \in N_{T'} \setminus S$.

2.5 Shapley Value

A *game* (with transferable utility) is a set function $v : 2^N \to \mathbb{R}$ such that $v(\varnothing) = 0$. In *Cooperative Game Theory* (CGT), N is the set of players, and $v(S)$ (for $S \subseteq N$) is the amount of wealth produced by S when they cooperate. The Shapley value is a fair share of the global wealth $v(N)$ produced by all players together, among themselves [Shapley, 1953]:

$$\phi_i^{\mathrm{Sh}}(N, v) := \sum_{S \subseteq N \setminus i} \frac{(n - |S| - 1)!|S|!}{n!} \big[v(S \cup \{i\}) - v(S)\big]. \tag{6}$$

It is characterized by four properties: *Additivity* ($\phi_i^{\mathrm{Sh}}(N, v + w) = \phi_i^{\mathrm{Sh}}(N, v) + \phi_i^{\mathrm{Sh}}(N, w)$), *Null player* (if $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N \setminus \{i\}$, then $\phi_i^{\mathrm{Sh}}(N, v) = 0$), *Symmetry* ($\phi_{\pi k}^{\mathrm{Sh}}(\pi N, \pi v) = \phi_k^{\mathrm{Sh}}(N, v)$ for every permutation π on N), and *Efficiency* ($\sum_{i \in N} \phi_i^{\mathrm{Sh}}(N, v) = v(N)$) [Shapley, 1953].

3 Related Works

In MCDA, complete explanations have been proposed for the additive utility model [Belahcene *et al.*, 2017] and the majority model [Labreuche *et al.*, 2012]. The explanation can also be obtained thanks to a surrogate model [Zhong *et al.*, 2014] or to explanation schema [Labreuche, 2011; Nunes *et al.*, 2014].

Sensitivity-Based Explanation and the Idea of Compellingness. Sensitivity analysis can be used as an explanation means, by computing the degree to which the parameters can be changed, while keeping untouched the decision. Alvarez applies this principle to decision trees, by identifying the list of the tests of the tree that are the most sensitive to a change in the input data [Alvarez, 2004]. In [Visser *et al.*, 2014], a hierarchical qualitative preference model is explained, by tracing from the top overall node down to elementary criteria, which element counts in the decision.

In the family of methods based sensitivity analysis, some authors have defined the contribution level of each criterion in some decision. For the weighted sum WS model, the *compellingness* is a measure of the contribution of criterion i in the differences of scores $\mathrm{WS}(a) - \mathrm{WS}(b)$ in the comparison of two options a and b in $[0, 1]^n$, with $\mathrm{compel}(i, a, b) := w_i \, |a_i - b_i|$ [Klein, 1994; Carenini and Moore, 2006].

The notion of compellingness has been extended to the Choquet integral: $\mathrm{compel}(i, a, b) = \left|a_i \, \Delta_i^a - b_i \, \Delta_i^b\right|$ [Montmain *et al.*, 2005]. The explanation then consists in displaying the K largest values of the compellingness, where K can be derived by several strategies [Klein, 1994; Carenini and Moore, 2006].

For the weighted sum, the compellingness is a relevant measure of the contribution of each criterion in the decision. On the other hand, the compellingness defined for a Choquet integral [Montmain *et al.*, 2005] has several drawbacks. First of all, if $a \in \Sigma_\sigma$ is close to being ordered by another order σ', this means that changing slightly a could yield a completely different explanation (thereby using σ' instead of σ). Hence the explanation is not continuous (stable) in a. **Continuity** is an important principle as the values of the alternatives on the attributes may be imprecise and affected by small errors. Secondly, focusing on only one ranking σ amounts to doing as if the Choquet integral (which can represent interaction among criteria) behaved linearly, as a simple weighted sum. Interaction is a global behavior that cannot be explained by only looking at the local weights. Interaction comes from the fact the local weights depend on the value of the scores. Lastly, the approach of [Montmain *et al.*, 2005] is designed for a specific aggregation function within a flat organization of the criteria, and therefore cannot apply for more general hierarchical models, such as the multi-linear model or the GAI model. The influence index we are trying to construct in this paper aims at being applicable to any utility-based MCDA model U (a Choquet integral, a GAI model, a weighted sum,...) defined on a hierarchy of criteria, and thus it should work directly on the utility function U rather on weights.

4 Axiomatic Properties

Due to the potential large number of criteria, the presence of interacting criteria and the hierarchical structure, we are aiming at an explanation emphasizing on the main contributing attributes. Given a tree T of criteria, we wish to construct an index $I_i(x, y, T, U) \in \mathbb{R}$ measuring the *influence* of attribute i in the difference of scores $U(y) - U(x)$ (comparison between alternatives x and y by model U) on tree T. When there is no confusion, it will be simply denoted by $I_i(x, y)$.

In Ex. 1, the OC might be interested in the influence of aggregation nodes such as "kinematics" (node 9) in the comparison of x and y. Hence I_i has to be defined for every $i \in M_T$. Moreover, the results of the explanation can be useful for different DMs, who have different roles and competences. In Ex. 1, the Provincial Governor (PG) might be involved to validate a decision on a specific suspicious ship. The PG is not interested in too many details and a simplified tree might be presented to them, such as the one in Fig. 2. If the subtree representing the concerns of a DM is $T' \in \mathcal{T}_T$, one shall thus also compute indices $I_i(x^{T'}, y^{T'}, T', U_{T'})$.

We denote by Δ_T the set of vectors (x', y', T', U') where $T' \in \mathcal{T}_T$, $x', y' \in X_{N_{T'}}$ and $U' \in \mathcal{U}_{T'}$. Index I is thus seen as a mapping from Δ_T onto a vector of real numbers, where, for $(x', y', T', U') \in \Delta_T$, $I(x', y', T', U') = \big(I_i(x', y', T', U')\big)_{i \in M_{T'}}$.

We adopt an axiomatization approach. As the axioms may not uniquely specify the influence, we consider a set $\mathcal{I}(\Delta_T)$ of such mappings I. The following axioms will be given on $\mathcal{I}(\Delta_T)$.

Property #1: Restricted Values.

It is difficult to identify which attribute is the most influential in the comparison between x and y, as, in general, x and y take different values on all attributes. In order to identify the most influential variables, one can analyze the consequence of going from x_i to y_i on attribute i – the values on the other

attributes being fixed. The move from x_i to y_i can be done in several ways: in only one move, with a finite number of intermediate steps, or in a continuous way. Yet the attributes can be discrete or continuous. Moreover, x_i and y_i can be the only values of X_i. In Ex. 1, all X_i have only two values. As we wish to use the same treatment in all cases, we consider the move from x_i to y_i which makes the fewest assumptions: directly from x_i to y_i with no intermediate point. Hence we do not use other values than x_i, y_i from X_i. As this happens for all attributes in N, the influence index I shall depend only on the overall utility of alternatives of the form (y_S, x_{-S}):

> **Restricted Values (RV):** For $(x, y, T', U) \in \Delta_T$ and $I \in \mathcal{I}(\Delta_T)$, index $I(x, y, T', U)$ depends only on $\{U(y_S, x_{-S})\ ,\ S \subseteq N_{T'}\}$.

Property #2: Null Attribute.

Attribute $i \in N_{T'}$ is said to be *null* for x, y, U when $U(y_{S\cup\{i\}}, x_{-(S\cup\{i\})}) = U(y_S, x_{-S})$ for all $S \subseteq N_{T'} \setminus \{i\}$. Then, changing value x_i to y_i on attribute i has no impact on the utility, whatever the value of the remaining attributes. In Ex. 3, comparing $x =$ (Yes,None,High,High,High,High) with $y =$ (Yes,High,High,Low,Low,Low), the second attribute is null (as attribute 3 takes value 'High'). Note that this comes from the redundancy between criteria 2 and 3 (see Ex. 3). Hence its influence should be zero.

> **Null Attribute (NA):** Let $(x, y, T', U) \in \Delta_T$ and $I \in \mathcal{I}(\Delta_T)$. If an attribute $i \in N_{T'}$ is null for x, y, U, then
>
> $$I_i(x, y, T', U) = 0. \tag{7}$$

Property #3: Restricted Equal Treatment.

For $k, l \in N_{T'}$, we say that $k \sim_{x,y,T',U} l$ when $U(y_{S\cup\{k\}}, x_{-(S\cup\{k\})}) = U(y_{S\cup\{l\}}, x_{-(S\cup\{l\})})$ for all $S \subseteq N_{T'} \setminus \{k, l\}$. This means that attribute k is as desirable as l, as they return the same utility if we switch the values between these two attributes, whatever the values on the remaining attributes. In the case of a flat organization of criteria, due to symmetry reasons, attributes k and l should have the same influence level if $k \sim_{x,y,T',U} l$. For a general tree structure T', this argument is still valid when attributes k and l belong to the same part in the tree, i.e. when they have the same parent in T' (i.e. there exists $j \in M_{T'}$ such that $k, l \in \mathrm{Ch}_{T'}(j)$). In Ex. 3, attributes 4 and 5 are treated symmetrically in H_8. Hence if $x_4 = y_4$ and $x_5 = y_5$, then $4 \sim_{x,y,T',U} 5$, and attributes 4 and 5 deserve to have the same influence. More generally, two equivalent attributes according to $\sim_{x,y,T',U}$ should have the same influence.

> **Restricted Equal Treatment (RET):** Let $(x, y, T', U) \in \Delta_T$ and $I \in \mathcal{I}(\Delta_T)$. If $k, l \in N_{T'}$ have the same parent in T' and $k \sim_{x,y,T',U} l$, then
>
> $$I_k(x, y, T', U) = I_l(x, y, T', U). \tag{8}$$

Property #4: Additivity.

Most utility models U have some decomposability property. For instance, the weighted sum, the Choquet integral and the GAI model are written as a sum of submodels focused on less attributes. One could use this additivity property to compute the influence index of U from the influence indices computed for each submodel. This would help to reduce the computation burden. The additivity property says that the influence for a sum of two utilities is equal to the sum of the influences for the two utilities.

> **Additivity (ADD):** Let $I \in \mathcal{I}(\Delta_T)$ and $(x, y, T', U), (x, y, T', U') \in \Delta_T$ such that $(x, y, T', U + U') \in \Delta_T$. Then
>
> $$I(x, y, T', U + U') = I(x, y, T', U) + I(x, y, T', U'). \tag{9}$$

Note: we have seen in Sect. 3 that the compellingness for the Choquet integral is not continuous in U, but also justified that continuity is an important property. In this respect, requirement for additivity is very convenient as it automatically ensures continuity.

Property #5: Generalized Efficiency.

We wish to explain the comparison between two options x and y. This shall cover the preference relation between x and y (x is preferred/indifferent/less preferred to y), but also the intensity of preference (e.g. x is slightly/significantly/extremely preferred to y). The intensity of preference is quantified by $U(y) - U(x)$. We wish to identify how $U(y) - U(x)$ is *shared* among the nodes and attributes in T. As the influence is defined for each node in T, the influence of the top node s_T is set by convention to $U(y) - U(x)$.

As for the efficiency property in CGT, $U(y) - U(x)$ shall be split among the attributes: $\sum_{i\in N_{T'}} I_i(x, y, T', U) = U(y) - U(x)$. To go one step further, the influence at an aggregation node $l \in M_T \setminus N_T$ shall be split into the influence degree of its direct contributors $\mathrm{Ch}_T(l)$. In Ex. 1, the influence of node 7 is the sum of the influences of nodes 2 and 3. More formally, we obtain the following property.

> **Generalized Efficiency (GE):** Let $(x, y, T', U) \in \Delta_T$ and $I \in \mathcal{I}(\Delta_T)$. Then
>
> $$I_{s_T}(x, y, T', U) = U(y) - U(x), \tag{10}$$
>
> $$\forall l \in M_{T'} \setminus N_{T'}, \quad \sum_{i\in \mathrm{Ch}_{T'}(l)} I_i(x, y, T', U) = I_l(x, y, T', U). \tag{11}$$

Example 5 (Ex. 3 cont.) *In the tree T' of Fig. 2, relations (10) and (11) give $I_{10}(x, y, T', U) = U(y) - U(x) = I_1(x, y, T', U) + I_7(x, y, T', U) + I_9(x, y, T', U)$, $I_7(x, y, T', U) = I_2(x, y, T', U) + I_3(x, y, T', U)$.* ■

Property #6: Consistency with Restricted Tree.

Set $\mathcal{T}_T$ describes different granularity levels on T where some parts of T are hidden. For $j \in M_T$, trees T and $T_{[j]}$ represent the right level of detail for two DMs (e.g. T is Fig. 1 for the OC and $T_{[j]}$ is Fig. 2 for the PG). There is no reason to present two different values of the influence index at a same node i to them. We enforce this condition only when i and j belong to different parts of the tree (i.e. $i \notin \mathrm{Desc}_T(j)$ and $j \notin \mathrm{Desc}_T(i)$). Hence we have the next axiom.

Consistency with Restricted Tree (CRT): Let $(x, y, T', U) \in \Delta_T$ and $I \in \mathcal{I}(\Delta_T)$. Let $i, j \in M_{T'}$ s.t. $i \notin \mathrm{Desc}_{T'}(j)$ and $j \notin \mathrm{Desc}_{T'}(i)$. Then

$$I_i(x, y, T', U) = I_i(x^{T'_{[j]}}, y^{T'_{[j]}}, T'_{[j]}, U_{T'_{[j]}}). \quad (12)$$

Example 6 (Ex. 4 cont.) *For T given by Fig. 1, $i = 1$ and $j = 9$, $1 \notin \mathrm{Desc}_T(9) = \{4, 5, 6, 8, 9\}$ and $9 \notin \mathrm{Desc}_T(1) = \{1\}$. $T_{[9]}$ is given in Fig. 2.* **CRT** *says that the influence index of node 1 can be computed indifferently from T (for the OC) and $T_{[9]}$ (for the PG), with the same result.* ■

5 Expression of the Influence Index

We propose a construction of an influence index, taking inspiration from the Shapley value. It is not easy to determine the contribution of each attribute in $U(y) - U(x)$, as x and y generally take different values on all attributes. In order to distinguish the contribution of each attribute, we move from x to y changing one attribute at a time, following an ordering π on N. We obtain the following sequence of options: x, $(y_{\{\pi(1)\}}, x_{-\{\pi(1)\}})$, $(y_{\{\pi(1),\pi(2)\}}, x_{-\{\pi(1),\pi(2)\}})$, …, y. The influence of attribute i in ordering π writes:

$$\delta_\pi^{x,y,T,U}(i) := U(y_{S_\pi(i)}, x_{-S_\pi(i)}) - U(y_{S_\pi(i)\setminus\{i\}}, x_{-S_\pi(i)\setminus\{i\}}), \quad (13)$$

where $S_\pi(\pi(k)) := \{\pi(1), \ldots, \pi(k)\}$. The influence index of attribute i – in the spirit of the Shapley value – is then simply the average value of $\delta_\pi^{x,y,T,U}(i)$ over all orderings π. Note that this corresponds exactly to the Shapley value ϕ_i^{Sh} of game $v_{x,y,T,U} : 2^N \to \mathbb{R}$ defined by $v_{x,y,T,U}(S) = U(y_S, x_{-S}) - U(x)$.

As noted by Owen, the Shapley value is not suited when players are organized in a coalition structure [Owen, 1977]. Instead of considering all permutations on N, Owen just considers the permutations preserving the coalition structure. Unfortunately, the Owen solution is only defined for coalition structures and not for general trees.

Generalizing the Owen approach, we define, for a general tree T, the set of admissible orderings $\Pi(T)$ of N as the set of orderings of elements of N for which all elements of a subtree of T are consecutive. More precisely, $\pi \in \Pi(T)$ iff, for every $l \in M_T \setminus N$, indices $\pi^{-1}(\mathrm{Leaf}_T(l))$ are consecutive. A permutation π will be assimilated to a vector of n components: $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$.

Example 7 (Ex. 1 cont.) *For T given by Fig. 1, $\Pi(T)$ contains, for instance, elements $(5, 4, 6, 2, 3, 1)$ (indicating that $\pi(1) = 5$, $\pi(2) = 4$, $\pi(3) = 6$, $\pi(4) = 2$, $\pi(5) = 3$, $\pi(6) = 1$), $(1, 6, 4, 5, 2, 3)$ and $(1, 2, 3, 4, 5, 6)$, which all correspond to a reorganization of the tree that keeps its structure. On the other hand, $\Pi(T)$ does not contain $(2, 3, 4, 5, 1, 6)$, since, for $l = 9$, the elements in $\pi^{-1}(\mathrm{Leaf}_T(9)) = \pi^{-1}(\{4, 5, 6\}) = \{3, 4, 6\}$ are not consecutive. Indeed, attribute 1 is interleaved between attributes $\{4, 5\}$ and $\{6\}$ in order π.* ■

Using $\Pi(T)$, the influence of attribute $i \in N_T$ in the tree T can be defined by (where EOw stands for *Extended Owen*)

$$I_i^{\mathrm{EOw}}(x, y, T, U) = \begin{cases} \frac{1}{|\Pi(T)|} \sum_{\pi \in \Pi(T)} \delta_\pi^{x,y,T,U}(i) \text{ if } i \in N_T \\ \sum_{k \in \mathrm{Leaf}_T(i)} I_k^{\mathrm{EOw}}(x, y, T, U) \text{ else.} \end{cases}$$

The second formula follows from General Efficiency **GE**. The new index subsumes to the Shapley value for flat organizations and to the Owen value for coalition structures.

When $U(y) > U(x)$, $I_i^{\mathrm{EOw}}(x, y, T, U) = 0$ (resp. > 0, and < 0) means that attribute i does not contribute (resp. contributes positively, and contributes negatively) on the preference of y over x by U.

The next result shows that I^{EOw} is the only solution fulfilling all previous axioms.

Theorem 1 *There is a unique influence index satisfying* **RV**, **NA**, **RET**, **ADD**, **GE** *and* **CRT**. *It is equal to I^{EOw}.*

The proofs of Theorems 1 through 5 (see below) are omitted due to space limitation.

This result generalizes an axiomatic characterization of the Owen value to coalition structures [Owen, 1977]. We note that axioms **NA**, **RET**, **ADD** and **E** restricted to coalition structures were used in [Owen, 1977], and Owen introduced a last axiom called *Intermediate Game Property*, which is quite similar to (11). Lastly **CRT** is completely new and is related to the hierarchical structure.

6 Computational Complexity Analysis

Index I_i^{EOw} has an exponential complexity, like the Shapley value, making its computation intractable, even for small values of n. We show that we can drastically speed-up the computation time, by taking profit of symmetries among permutations in $\Pi(T)$. This idea was proposed by Owen for his value [Owen, 1977]. We extend it to any tree.

The path from s_T to i in T consists of the nodes $r_0 = s_T$, r_1, …, $r_t = i$. Iterating over **CRT**, the next result is shown.

Theorem 2 *Let $J = \bigcup_{l=1}^{t-1} \mathrm{Ch}_T(r_{l-1}) \setminus \{r_l\}$. We have*

$$I_i^{\mathrm{EOw}}(x, y, T, U) = I_i^{\mathrm{EOw}}(x^{T_{[J]}}, y^{T_{[J]}}, T_{[J]}, U_{T_{[J]}}). \quad (14)$$

The extended Owen value of node i for tree T can be computed equivalently on tree $T_{[J]}$.

Example 8 (Ex. 4 cont.) *For T given by Fig. 1, and $i = 2$, we obtain $J = \{1, 3, 9\}$. Hence the influence of node 2 can be computed from $T_{[J]}$ given by Fig. 2.* ■

The computational complexity is still exponential but on a much smaller tree. More precisely, the right hand side of (14) requires $C(i, T) := 2 \prod_{l=1}^{t} 2^{|\mathrm{Ch}_T(r_{l-1})|-1}$ computations of an overall utility U .

We now show that this is sufficient to compute $I_i^{\mathrm{EOw}}(x, y, T, U)$ in reasonable time. Most real-life MCDA models contain less than 40 criteria, aggregate at most 6 nodes at each level and have maximal depth of 4. We randomly generated trees having at most 100 criteria, at most 6 children at each level and maximal depth of 5. We also randomly generated options x, y and 2-additive Choquet integrals for U. For each instance, we store the average time to compute one index I_i^{EOw} (average over $i \in N_T$). Table 1 shows the computation times over 25 000 generations performed on a computer equipped with 3.1 GHz Intel Core i7.

The computations are fast and compatible with user interaction up to 40 attributes. The time increases but remains

n	$Q_{1/20}$	$Q_{1/4}$	$Q_{1/2}$	$Q_{3/4}$	$Q_{19/20}$
2-10	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.0006
11-20	$< 10^{-5}$	0.0009	0.002	0.004	0.022
21-30	0.0014	0.004	0.009	0.03	0.18
31-40	0.005	0.014	0.035	0.102	0.731
41-50	0.0115	0.0337	0.0874	0.249	1.545
51-60	0.0219	0.0652	0.183	0.538	3.882
61-70	0.0376	0.121	0.316	0.948	5.484
71-80	0.0697	0.225	0.561	1.537	7.619
81-90	0.106	0.326	0.798	2.402	12.411
91-100	0.178	0.542	1.239	3.148	16.336

Table 1: Different quantiles of the distribution of the average time to compute one index I_i^{EOw}, in seconds for each interval of n. For $\tau \in [0,1]$, Q_τ is the value of the distribution for which the set of elements lower than Q_τ represents exactly a proportion τ of all the elements. In particular, $Q_{1/2}$ is the median.

admissible for most of larger trees. We analyzed the structure of trees for which the computation time is quite large. They correspond to ill-balanced trees for which $T_{[J]}$ is not so different from T. Such trees are very unlikely to arise from a real-life MCDA problem. We have tested the computation time on several real-life problems. For instance, for a real-life model with 29 criteria, the time to compute one index, with and without truncation, is equal to 0.05s and 0.8s respectively.

When complexity $C(i,T)$ is beyond a given threshold, one may consider that the computation time is not admissible for the user. It is then possible to cut the tree T at a given depth in such a way that $C(i,T')$ remains below the threshold, where T' is the cut tree. For example, cutting the tree of Fig. 1 at depth 2 means that nodes 4 and 5 are removed. Doing this means that the influence of deep nodes (not in the cut tree) are not presented to the user, which is admissible for very large trees. It is easy to find the largest cut depth fulfilling the previous requirement.

The number of alternatives is generally small in MCDA. When there is an exponential number of options, we hopefully do not need to explain all comparisons. For instance, in incremental elicitation using minmax regret [Boutilier *et al.*, 2006], one only needs to explain the comparison between optimal minmax regret solution and its worst contester.

7 How to Answer to the Needs N1, N2 and N3?

This section describes how the influence index can serve to answer to the three needs **N1**, **N2** and **N3** (see Sect. 1).

7.1 Case of Need N1 (Interpretability)

It is not easy to provide a general interpretation of a hierarchical model U. This task can be performed thanks to our new indicator. The mean importance of attribute i is simply the influence index of comparing the two alternatives $x^\perp$ and $x^\top$ having the smallest and largest possible utilities on N respectively (i.e. $u_i(x_i^\perp) = 0$ and $u_i(x_i^\top) = 1$ for all i), as this covers the range of X_i's. The next result shows that, for the 2-additive Choquet integral, the computation becomes easy.

Theorem 3 *If all aggregation functions are* 2*-additive Choquet integral, then*

$$I_i^{\mathrm{EOw}}(x^\perp, x^\top, T_{[J]}, U) = \prod_{l=1}^{t} \mathrm{Imp}_{r_l}(V_l, m^l). \qquad (15)$$

where m^l are the Möbius coefficients at aggregation node r_{l-1} (for $l \in \{1,\ldots,t\}$) in the path from s_T to i.

Crit	1	2	3	4	5	6	U
$x^\perp$	No	None	None	Low	Low	Low	0
$x^\top$	Yes	High	High	High	High	High	1
x'	Yes	None	None	High	High	High	$\frac{2}{3}$
x''	No	High	High	Low	Low	Low	$\frac{1}{3}$
x'''	Yes	None	None	Low	Low	Low	$\frac{1}{3}$

Table 2: Performance values

I_i^{EOw}	10	1	7	2	3	9	8	4	5	6
$x^\perp, x^\top$	1	1/3	1/3	2/15	1/5	1/3	1/6	1/12	1/12	1/6
$x^\perp, x'$	2/3	1/2	0	0	0	1/6	1/12	1/24	1/24	1/12
$x'', x^\top$	2/3	1/6	0	0	0	1/2	1/4	1/8	1/8	1/4
$x''', x^\top$	$\frac{2}{3}$	0	1/6	1/15	1/10	1/2	1/4	1/8	1/8	1/4

Table 3: Influence indices for all nodes in M_T

If all aggregation functions are weighted sums, we would expect that the importance of criterion i is the product of the weights of node i and its parents in the path from leaf i to the root, which is exactly what Theorem 3 gives. Formula (15) nicely generalizes this expression, as the importance of an attribute i is the product of the relative importance of this node and its parents in the path from leaf i to the root. For a flat organization, we also recover that the mean importance of attribute i is $\mathrm{Imp}_{r_l}(V_l, m^l)$ (see Sect. 2.3). Theorem 3 shows that our influence index, despite its combinatorial expression, subsumes to a very simple and intuitive expression in the case where x and y are the extreme values. We now illustrate this on the running example on $x^\perp, x^\top$ given by Table 2.

Example 9 (Ex. 3 cont.) *$I_i^{\mathrm{EOw}}(x^\perp, x^\top)$ is given by Table 3. We see that the three top level nodes $1, 7, 9$ have the same importance, as expected from the expression of v_{10}^U. On the contrary, the importance of node 7 is unequally shared between nodes 2 and 3, in coherence with the expression of v_7^U.*

7.2 Case of Need N2 (Explicability)

There is no competitive explanation approach for hierarchical MCDA models. In order to compare I^{EOw} with the existing works, we restrict ourselves to flat organizations.

Weighted Sum. Consider a weighted sum on a flat organization of the criteria: $U(x) = \sum_{i \in N_T} w_i\ u_i(x_i)$. The compellingness is then $\mathrm{compel}(i, u(x), u(y)) = w_i\,|u_i(y_i) - u_i(x_i)|$, where $u(x) := (u_1(x_1), \ldots, u_n(x_n))$.

Theorem 4 *For a weighted sum, we have*

$$I_i^{\mathrm{EOw}}(x, y, T, U) = w_i\,(u_i(y_i) - u_i(x_i)). \qquad (16)$$

Our influence index differs from the compellingness only by the absolute value operator. We prefer to keep the sign in order to keep the sense of the preference between x and y.

Two-Additive Choquet Integral. Consider a flat organization of the criteria where the aggregation function is a 2-additive Choquet integral w.r.t. Möbius coefficients m. The compellingness is then $\mathrm{compel}(i, u(x), u(y)) = \left| u_i(x_i)\, \Delta_i^{u(x)} - u_i(y_i)\, \Delta_i^{u(y)} \right|$.

The next result gives the analytical expression of I_i^{EOw} for a 2-additive Choquet integral with Möbius coefficients m.

Theorem 5 *We have*

$$I_i^{\mathrm{EOw}}(x, y, T, U) = m_i\,(u_i(y_i) - u_i(x_i)) + \sum_{k \in N \setminus \{i\}} m_{i,k}\, D_{i,k}(x, y) \quad (17)$$

where $D_{i,k}(x, y) = \frac{d_{i,k}(x_i, y_i, x_k) + d_{i,k}(x_i, y_i, y_k)}{2}$ *and* $d_{i,k}(x_i, y_i, z_k) = u_i(y_i) \wedge u_k(z_k) - u_i(x_i) \wedge u_k(z_k)$.

The next example compares our solution with compel.

Example 10 *Let us consider* $U(z_1, z_2) = \frac{u_1(z_1) + u_1(z_1) \wedge u_2(z_2)}{2}$ *and the two options* x, y *with* $u(x) = (0.3, 0.9)$ *and* $u(y) = (1, 0.9)$. *We have* $U(x) = 0.3$ *and* $U(y) = 0.95$. *The two solutions return completely different results:* $\mathrm{compel}(1, u(x), u(y)) = 0.2$, $\mathrm{compel}(2, u(x), u(y)) = 0.45$, $I_1^{\mathrm{EOw}}(x, y) = 0.65$ *and* $I_2^{\mathrm{EOw}}(x, y) = 0$. *The ordering among the influences of the two attributes is even different. As attribute* 2 *is null* ($u_2(x_2) = u_2(y_2)$), *the influence of attribute* 2 *should be* 0 *(see* **NA**), *as obtained by* I^{EOw}. *It is thus very unnatural that* $\mathrm{compel}(2, u(x), u(y))$ *returns a much larger value than that for criterion* 1.

Consider now two other options x', y' *close to* x, y: $u(x') = (0.3, 1)$ *and* $u(y') = (0.9, 1)$. *The two solutions now return the same result:* $\mathrm{compel}(1, u(x), u(y)) = I_1^{\mathrm{EOw}}(x, y) = 0.6$, $\mathrm{compel}(2, u(x), u(y)) = I_2^{\mathrm{EOw}}(x, y) = 0$. *We note that* I^{EOw} *gives very stable results between* x, y *and* x', y'. *On the contrary, the values of* compel *are completely different for* x, y *and for* x', y', *which yields stability problems.*

We now illustrate the explanation on the running example.

Example 11 (Ex. 3 cont.) *The maritime patrol is interested in the way the PL of a ship varies over time. Consider a ship whose state evolves from state* $x^\perp$ *to* x', *which implies a very significant increase of PL (from* 0 *to* $\frac{2}{3}$*). How to explain this? Computation in table 3 shows that* $I_1^{\mathrm{EOw}}(x^\perp, x') = \frac{1}{2}$ *and* $I_9^{\mathrm{EOw}}(x^\perp, x') = \frac{1}{6}$*: when there is a 'low risk' (no suspicion of illegal activity, as* $x_2^\perp = x_3^\perp = x_2' = x_3' = \mathrm{None}$*), 'AIS/radar incoherence' (node* 1*) has a larger impact on PL than having a fast boat approaching shore. Indeed, as* $v_7^U(x^\perp) = v_7^U(x') = 0$, *any improvement of* v_9^U *(resp.* v_1^U*) has an impact on only the second term in* v_{10}^U *(resp. on the first two terms in* v_{10}^U*). Hence the values of* I^{EOw} *are natural. In situation of stress, the OC is asking for very simple information. The system can tell to the user that the increase of PL between* $x^\perp$ *to* x' *mainly comes from node* 1. *The difference of influence indices between nodes* 1 *and* 9 *is all the more noticeable as the mean importance of nodes* 1 *and* 9 *are similar* ($I_1^{\mathrm{EOw}}(x^\perp, x^\top) = I_9^{\mathrm{EOw}}(x^\perp, x^\top)$).

Let us consider another ship with state changing from x'' *to* $x^\top$ *(see Table 2). Note that the attribute changes from* x'' *to* $x^\top$ *are similar to that between* $x^\perp$ *and* x'. *However, the explanation of the difference between* x'' *and* $x^\top$ *is completely different, as it now comes from the 'Capability to escape', critical during a 'high risk' situation (suspicion of illegal activity, as* $v_7^U(x'') = v_7^U(x^\top) = 1$*), since information of 'AIS/radar incoherence' is redundant with suspicion of illegal activity:* $I_1^{\mathrm{EOw}}(x'', x^\top) = \frac{1}{6}$ *and* $I_9^{\mathrm{EOw}}(x'', x^\top) = \frac{1}{2}$. *The complete change of explanation in the previous two comparisons accurately reflects the model behavior and in particular the strong interactions present between the criteria.*

7.3 Case of Need N3 (Sensibility Analysis)

We illustrate **N3** on the running example.

Example 12 (Ex. 3 cont.) *The OC now focuses their attention on ship* x''' *having AIS inconsistency (Att.* 1*). The values of attributes* $2, 3, 4, 5$ *are unknown and set to default None/Low value. While it is crucial for the OC to determine true values of these attributes, it is too time consuming to investigate on all of them. On which attributes shall the OC focus their attention? This can be obtained from the computation of the influence degree between* x''' *and the most threatening situation* $x^\top$. *It is not so important to investigate on smuggling (att.* 2 *and* 3*) as their influence is low. The maritime patrol puts more means to identify whether the ship is a fast boat, as attributes* 4 *and* 5 *have a large influence.*

8 Conclusion and Perspectives

We have defined an influence index of a MCDA model for the purpose of explanation, from two approaches: an axiomatic approach which brings a formal justification, and a constructive approach which provides a simple interpretation. We also show how to drastically reduce the complexity of its computation, making it usable in practical applications.

Our approach has several benefits. First of all, it can be applied to any quantitative MCDA model (computing a utility) on any hierarchy of criteria, whereas the existing approaches are dedicated to a specific model and are restricted to a flat organization of criteria. Moreover, it is simple enough to be employed by end-users, and without prior knowledge in MCDA. Finally it can be computed at any node in the tree of criteria and aggregation nodes, allowing the user to perform an analysis at any level.

For future work, we will aim at further reducing the computation time for particular expressions of U (e.g. for Choquet integrals), and testing the index on real users.

Apart from the new explanation approach, we introduced a new Shapley value on graphs. This value could also be used in argumentation [Amgoud *et al.*, 2017] or inconsistencies in databases [Hunter and Konieczny, 2010].

Acknowledgments

This work has been supported by the European project H2020/CIP-2016-2017-1/740898, DEFENDER "Defending the European Energy Infrastructures ".

References

[Alvarez, 2004] I. Alvarez. Explaining the result of a decision tree to the end-user. In *16th European Conference on Artificial Intelligence (ECAI'2004)*, pages 411–415, Valencia, Spain, 2004.

[Amgoud *et al.*, 2017] L. Amgoud, J. Ben-Naim, and S. Vesic. Measuring the intensity of attacks in argumentation graphs with shapley value. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 63–69, Melbourne, Australia, August 2017.

[Belahcene *et al.*, 2017] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82:151–183, 2017.

[Boutilier *et al.*, 2006] C. Boutilier, R. Patrascu, P. Poupart, and D. Schuurmans. Constraint-based optimization and utility elicitation using the minimax decision criterion. *Artificial Intelligence*, 170(8-9):686–713, 2006.

[Carenini and Moore, 2006] G. Carenini and J.D. Moore. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170:925–952, 2006.

[Choquet, 1953] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953.

[Datta *et al.*, 2016] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence. In *IEEE Symposium on Security and Privacy*, San Jose, CA, USA, May 2016.

[Diestel, 2005] R. Diestel. *Graph Theory*. Springer-Verlag, New York, 2005.

[Fishburn, 1967] P. Fishburn. Interdependence and additivity in multivariate, unidimensional expected utility theory. *International Economic Review*, 8:335–342, 1967.

[Grabisch, 1996] M. Grabisch. The application of fuzzy integrals in multicriteria decision making. *European J. of Operational Research*, 89:445–456, 1996.

[Grabisch, 1997] M. Grabisch. Alternative representations of discrete fuzzy measures for decision making. *Int. J. of Uncertainty, Fuzziness, and Knowledge Based Systems*, 5:587–607, 1997.

[Hunter and Konieczny, 2010] A. Hunter and S Konieczny. On the measure of conflicts: Shapley inconsistency values. *Artificial Intelligence*, 174:1007–1026, 2010.

[Keeney and Raiffa, 1976] R. L. Keeney and H. Raiffa. *Decision with Multiple Objectives*. Wiley, New York, 1976.

[Klein, 1994] D.A. Klein. *Decision analytic intelligent systems: automated explanation and knowledge acquisition*. Lawrence Erlbaum Associates, 1994.

[Labreuche *et al.*, 2012] Ch. Labreuche, N. Maudet, and W. Ouerdane. Justifying dominating options when preferential information is incomplete. In *European Conference on Artificial Intelligence (ECAI)*, Montepellier, France, August 27-31 2012.

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

[Lundberg and Lee, 2017] S. Lundberg and S.I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS) 30*, pages 4768–4777. Curran Associates, Inc., 2017.

[Miller, 1956] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

[Miller, 2017] T. Miller. Explainable AI: Insights from the social sciences. *ArXiv e-prints*, 1706.07269, https://arxiv.org/abs/1706.07269, 2017.

[Montmain *et al.*, 2005] J. Montmain, G. Mauris, and A. Akharraz. Elucidation and decisional risk in a multi criteria decision based on a Choquet integral aggregation: A cybernetic framework. *International Journal of Multi-Criteria Decision Analysis*, 13:239–258, 2005.

[Nunes *et al.*, 2014] I. Nunes, S. Miles, M. Luck, S. Barbosa, and C. Lucena. Pattern-based explanation for automated decisions. In *European Conference on Artificial Intelligence (ECAI)*, pages 669–674, Prague, Czech Republic, August 2014.

[Owen, 1977] G. Owen. Values of games with a priori unions. In O. Moeschlin R. Hein, editor, *Essays in Mathematical Economics and Game Theory*, pages 76–88. Springer Verlag, 1977.

[Ribeiro *et al.*, 2016] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, California, USA, 2016.

[Shapley, 1953] L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games, Vol. II*, number 28 in Annals of Mathematics Studies, pages 307–317. Princeton University Press, 1953.

[Visser *et al.*, 2014] W. Visser, K. Hindriks, and C. Jonker. Explaining qualitative preference models. In *6th Multidisciplinary Workshop on Advances in Preference Handling, European Conference on Artificial Intelligence*, Montpellier, France, 2014.

[Zhong *et al.*, 2014] Q. Zhong, X. Fan, F. Toni, and X. Luo. Explaining best decisions via argumentation. In *Proceedings of the European Conference on Social Intelligence (ECSI-2014)*, pages 224–237, Barcelona, Spain, November 2014.