Preferences and Ethical Principles in Decision Making

Andrea Loreggia
University of Padova
andrea.loreggia@gmail.com

Nicholas Mattei
IBM Research
n.mattei@ibm.com

Francesca Rossi
IBM Research
University of Padova
francesca.rossi2@ibm.com

K. Brent Venable
Tulane University
kvenabl@tulane.edu

Abstract

If we want AI systems to make decisions, or to support humans in making them, we need to make sure they are aware of the ethical principles that are involved in such decisions, so they can guide towards decisions that are conform to the ethical principles. Complex decisions that we make on a daily basis are based on our own subjective preferences over the possible options. In this respect, the CP-net formalism is a convenient and expressive way to model preferences over decisions with multiple features. However, often the subjective preferences of the decision makers may need to be checked against exogenous priorities such as those provided by ethical principles, feasibility constraints, or safety regulations. Hence, it is essential to have principled ways to evaluate if preferences are compatible with such priorities. To do this, we describe also such priorities via CP-nets and we define a notion of distance between the ordering induced by two CP-nets. We also provide tractable approximation algorithms for computing the distance and we define a procedure that uses the distance to check if the preferences are *close enough* to the ethical principles. We then provide an experimental evaluation showing that the quality of the decision with respect to the subjective preferences does not significantly degrade when conforming to the ethical principles.

Introduction

If we want people to trust AI systems, we need to provide them with the ability to discriminate between good and bad decisions. The quality of a decision should not be based only on the preferences or optimisation criteria of the decision makers, but also on other properties related to the impact of the decision, such as whether it is ethical, or if it complies to constraints and priorities given by feasibility constraints or safety regulations.

A lot of work has been done to understand how to model and reason with subjective preferences. This is understandable, since preferences are ubiquitous in everyday life. We use our own subjective preferences whenever we want to make a decision to choose our most preferred alternative. Therefore the study of preferences in computer science and AI has been very active for a number of years with important theoretical and practical results (Domshlak *et al.* 2011; Pigozzi *et al.* 2015) as well as libraries and datasets (Mattei and Walsh 2013).

Our preferences may apply to one or more of the individual components, rather than to an entire decision. For example, if we need to choose a car, we may prefer certain colours over others, and we may prefer certain brands over others. We may also have conditional preferences, such as in preferring red cars if the car is a convertible. For these scenarios, the CP-net formalism (Boutilier *et al.* 2004) is a convenient and expressive way to model preferences (Rossi *et al.* 2011; Chevaleyre *et al.* 2008; Goldsmith *et al.* 2008; Cornelio *et al.* 2013) CP-nets indeed provide an effective compact way to qualitatively model preferences over outcomes (that is, decisions) with a combinatorial structure. CP-nets are also easy to elicit and provide efficient optimization reasoning (Chevaleyre *et al.* 2011; Allen *et al.* 2015). Moreover, in a collective decision making scenario, several CP-nets can be aggregated, e.g., using voting rules (Conitzer *et al.* 2011; Mattei *et al.* 2013; Cornelio *et al.* 2015), to find compromises and reach consensus among decision makers.

If ethical constraints are added to this scenario, it means that the subjective preferences of the decision makers is not the only source of information we should consider (Sen 1974; Thomson 1985; Bonnefon *et al.* 2016). Indeed, depending on the context, we may have to consider specific ethical principles derived from an appropriate ethical theory (Copp 2005). While preferences are important, when preferences and ethical principles are in conflict, the principles should override the subjective preferences of the decision maker. For example, in a hiring scenario, the preferences of the hiring committee members over the candidates should be measured against ethical guidelines and laws e.g., ensuring gender and minority diversity. Therefore, it is essential to have principled ways to evaluate if preferences are compatible with a set of ethical principles, and to measure how much these preferences deviate from the ethical principles. The ability to precisely quantify the distance between subjective preferences and external priorities, such as those given by ethical principles, provides a way to both recognize deviations from feasibility or ethical constraints, and also to suggest more compliant decisions.

In this paper we use CP-nets to model both exogenous priorities, e.g., those provided by ethical principles, and subjective preferences of decision makers. Thus the distance between an individual subjective preferences and some ethical principles can be measured via a notion of distance between CP-nets. Indeed, we define such a notion of distance (formally a distance function or metric) between CP-nets. A more comprehensive discussion of CP-nets and distances between them is given by Loreggia *et al.* (2018).

Since CP-nets are a compact representation of a partial order over the possible decisions, the ideal notion of distance is a distance between the induced partial orders of the CP-nets. However, the size of the induced orders is exponential in the size of the CP-net, and we conjecture that computing a distance between such partial orders is computationally intractable because of this possibly exponential explosion. Therefore we propose a tractable approximation that is computed directly over the CP-nets dependency graphs, and we study the quality of the approximation.

To define the desired distance between partial orders, we generalize the classic (Kendall 1938) τ (KT) distance, which counts the number of inverted pairs between two complete, strict linear orders. We add a penalty parameter p defined for partial rankings as proposed by (Fagin *et al.* 2006), and use this distance, that we call KTD, to compare partial orders. In KTD the contribution of pairs of outcomes that are ordered in opposite ways is 1 and that of those that are ordered in one partial order and incomparable in the other is p. We show that $0.5 \leq p < 1$ is required for KTD to be a distance.

For the tractable approximation of KTD, we can define a distance between CP-nets, called CPD, that only analyzes the dependency structure of the CP-nets and their CP-tables. We then characterize the case when $CPD = 0$, which correspond to when the two CP-nets have the same dependency structure and CP-tables. In other words, $CPD = 0$ if and only if the two CP-nets are identical and they induce the same partial order over outcomes.

In general the values returned by CPD and KPD can be different. More precisely, the pairs of outcomes for which CPD could give an incorrect contribution to the distance are those that are either incomparable in both CP-nets (in this case CPD could generate an error of $+p$ or $-p$), or that are incomparable in a CP-net and ordered in the other (in this case the CPD error can be +1). To give upper and lower bounds to the error that CPD can make, we study the number of incomparable pairs present in a CP-net. We show that it is polynomial to compute the number of incomparable pairs of outcomes in a separable CP-net (that is, CP-nets with no dependencies among features). Non-separable CP-nets have fewer incomparable pairs of outcomes, since each dependency link eliminates at least one incomparable pair.

Our theoretical bounds are fairly wide. For this reason, we perform an experimental analysis of the relationship between CPD and KTD, which shows that the average error is never more than 10%. We then define a procedure that evaluates the distance between subjective preferences and ethical principles, and makes decisions using the subjective preferences if they are *close enough* to the ethical principles. Otherwise, the procedure moves to less preferred decisions until we find one that is a compromise between the ethical principles and the preferences. We then perform an experimental evaluation showing that the quality of the decision with respect to the subjective preferences does not significantly degrade, i.e., only needs to be moved a short distance in the preference order, when we need compliance with the ethical principles.

Background: CP-nets

CP-nets (Boutilier *et al.* 2004) (for Conditional Preference networks) are a graphical model for compactly representing conditional and qualitative preference relations. They are sets of *ceteris paribus* preference statements (cp-statements). For instance, the cp-statement *"I prefer red wine to white wine if meat is served."* asserts that, given two meals that differ *only* in the kind of wine served *and* both containing meat, the meal with red wine is preferable to the meal with white wine. Formally, a CP-net has a set of features $F = \{x_1, \ldots, x_n\}$ with finite domains $\mathcal{D}(x_1), \ldots, \mathcal{D}(x_n)$. For each feature x_i, we are given a set of *parent* features $Pa(x_i)$ that can affect the preferences over the values of x_i. This defines a *dependency graph* in which each node x_i has $Pa(x_i)$ as its immediate predecessors. An *acyclic* CP-net is one in which the dependency graph is acyclic. Given this structural information, one needs to specify the preference over the values of each variable x for *each complete assignment* on $Pa(x)$. This preference is assumed to take the form of a total or partial order over $\mathcal{D}(x)$. A cp-statement has the general form $x_1 = v_1, \ldots, x_n = v_n : x = a_1 \succ \ldots \succ x = a_m$, where $Pa(x) = \{x_1, \ldots, x_n\}$, $D(x) = \{a_1, \ldots, a_m\}$, and $\succ$ is a total order over such a domain. The set of cp-statements regarding a certain variable X is called the cp-table for X.

Consider a CP-net whose features are A, B, C, and D, with binary domains containing f and $\overline{f}$ if F is the name of the feature, and with the cp-statements as follows: $a \succ \overline{a}$, $b \succ \overline{b}$, $(a \wedge b) : c \succ \overline{c}$, $(\overline{a} \wedge \overline{b}) : c \succ \overline{c}$, $(a \wedge \overline{b}) : \overline{c} \succ c$, $(\overline{a} \wedge b) : \overline{c} \succ c$, $c : d \succ \overline{d}$, $\overline{c} : \overline{d} \succ d$. Here, statement $a \succ \overline{a}$ represents the unconditional preference for $A = a$ over $A = \overline{a}$, while statement $c : d \succ \overline{d}$ states that $D = d$ is preferred to $D = \overline{d}$, given that $C = c$.

A *worsening flip* is a change in the value of a variable to a less preferred value according to the cp-statement for that variable. For example, in the CP-net above, passing from $abcd$ to $ab\overline{c}d$ is a worsening flip since c is better than $\overline{c}$ given a and b. One outcome α is *better* than another outcome β (written $\alpha \succ \beta$) if and only if there is a chain of worsening flips from α to β. This definition induces a preorder over the outcomes, which is a partial order if the CP-net is acyclic.

Finding the optimal outcome of a CP-net is NP-hard (Boutilier *et al.* 2004). However, in acyclic CP-nets, there is only one optimal outcome and this can be found in linear time by sweeping through the CP-net, assigning the most preferred values in the cp-tables. For instance, in the CP-net above, we would choose $A = a$ and $B = b$, then $C = c$, and then $D = d$. In the general case, the optimal outcomes coincide with the solutions of a set of constraints obtained replacing each cp-statement with a constraint (Brafman and Dimopoulos 2004): from the cp-statement $x_1 = v_1, \ldots, x_n = v_n : x = a_1 \succ \ldots \succ x = a_m$ we get the constraint $v_1, \ldots, v_n \Rightarrow a_1$. For example, the following cp-statement (of the example above) $(a \wedge b) : c \succ \bar{c}$ would be replaced by the constraint $(a \wedge b) \Rightarrow c$.

In this paper we want to compare CP-nets while leveraging the compactness of the representation. To do this, we consider profile (P, O), where $\boldsymbol{P}$ is a collection of n CP-nets (whose graph is a directed acyclic graph (DAG)) over m common variables with binary domains and $\boldsymbol{O}$ is a total order over these variables. We require that the profile is O-legal (Lang and Xia 2009), which means that in each CP-net, each variable is independent to all the others following in the ordering O. Given a variable X_i the function $flw(X_i)$ returns the number of variables following X_i in O.

Since every acyclic CP-net is satisfiable (Boutilier *et al.* 2004), we compute a distance among two CP-nets by comparing a linearization of the partial orders induced by the two CP-nets. In this paper, we consider the linearization generated using the algorithm described in the proof of Theorem 1 of (Boutilier *et al.* 2004) and reproduced below as Algorithm 1. This algorithm works as follows: Given an acyclic CP-net A over n variables and a ordering O to which the A is O-legal, we know there is at least one variable with no parents. If more than one variable has no parents, then we choose the one that comes first in the provided ordering O; let X be such a variable. Let $x_1 \succ x_2$ be the ordering over $Dom(X)$ dictated by the cp-table of X. For each $x_i \in Dom(X)$, construct a CP-net, N_i, with the $n-1$ variables $V - X$ by removing X from the initial CP-net, and for each variable Y that is a child of X, revising its CPT by restricting each row to $X = x_i$. We can construct a preference ordering $\succ_i$ for each of the reduced CP-nets N_i. For each N_i recursively identify the variable X_i with no parents and construct a CP-net for each value in $Dom(X_i)$ following the same algorithm until a CP-net have variables. We can now construct a preference ordering for the original network A by ranking every outcome with $X = x_i$ as preferred to any outcome with $X = x_j$ if $x_i \succ x_j$ in CPT(X). This linearization, which we denote with $LexO(A)$, assures that ordered pairs in the induced partial order are ordered the same in the linearization and that incomparable pairs are linearized using the cp-tables.

Algorithm 1 Linearization of a Partial Order induced by a CP-net A

```
function LEXO($A, O, Lin = [], o = None$)    ▷ Where $A$ is a CP-net, $O$ is the O-legal order on $A$, $Lin$ is the (initially empty) linearization computed by the function, and $o$ is an outcome (initially none).
    if $O = Null$ then
        $Lin.append(o)$
        return $Lin$
    end if
    $v = pop(O)$
    for $value \in CPT_{A,o}(v)$ do
        $temp = o + value$
        $Lin = LexO(A, O, Lin, temp)$
    end for
    return $Lin$
end function
```

In Algorithm 1, $CPT_{A,o}(v)$ returns the ordered values of variable v in CP-net A, given a partial assignment o to a subset of variables. This linearization, which we denote with $LexO(A, O)$, where A is a CP-net and O an O-legal order over the features of A, enforces that ordered pairs in the induced partial order are ordered the same in the linearization and that incomparable pairs are linearized using the cp-tables.

A CP-net Distance Function

In what follows we will assume that all CP-nets are acyclic and in minimal (non-degenerate) form, i.e., all arcs in the dependency graph have a real dependency expressed in the cp-statements, see the extended discussion in (Allen *et al.* 2017; 2016). The following definition is an extension of the (Kendall 1938) τ (KT) distance with a penalty parameter p defined for partial rankings by (Fagin *et al.* 2006).

Definition 1. *Given two CP-nets A and B inducing partial orders P and Q over the same set of outcomes U:*

$$KTD(A, B) = KT(P, Q) = \sum_{\forall i,j \in U, i \neq j} K^p_{i,j}(P, Q) \quad (1)$$

where i and j are two outcomes with $i \neq j$, we have:

1. $K^p_{i,j}(P, Q) = 0$ *if i, j are ordered in the same way or they are incomparable in both P and Q;*
2. $K^p_{i,j}(P, Q) = 1$ *if i, j are ordered inversely in P and Q;*
3. $K^p_{i,j}(P, Q) = p$, $0.5 \leq p < 1$ *if i, j are ordered in P (resp. Q) and incomparable in Q (resp. P).*

In the previous definition we choose $p \geq 0.5$ to make $KTD(A, B)$ a distance function, indeed if $p < 0.5$ the distance does not satisfy the triangle inequality. We also exclude $p = 1$ so that there is a penalty for two outcomes being considered incomparable in one and ordered in another CP-net. This allows us, assuming O-legality, to define for each CP-net a unique most distant CP-net.

Proposition 1. *Given two acyclic CP-nets A and B that are not O-legal, deciding if $KTD(A, B) = 0$ cannot be computed in polynomial time unless $P = NP$.*

The NP-complete problem of checking for equivalence for two arbitrary CP-nets (Santhanam *et al.* 2013), i.e., deciding if two CP-nets induce the same ordering, can be reduced to the problem of checking if their KTD distance is 0. That is, if we had a polynomial time algorithm for deciding if $KTD(A, B) = 0$ then we could decide the equivalence problem for acyclic CP-nets. We know from (Boutilier *et al.* 2004) that dominance testing for max-δ-connected CP-nets, that is CP-nets where the maximum number of paths between two variables is polynomially bounded in the size of the CP-net is NP-complete. We know that O-legal, acyclic CP-nets are a class of max-δ-connected CP-nets because the O-legality constraint means that there are only a maximum of $n - 2$ paths between two nodes. However, this does not necessarily mean that the equivalence question is automatically hard. As we will see, our lower bound can actually be used to check equivalence for acyclic, O-legal CP-nets.

Since the question of dominance is closely related to that of distance, the complexity of computing KTD for O-legal CP-nets remains an important open question that we conjecture to be intractable. Due to this likely intractability we will define another distance for CP-nets which can be computed efficiently directly from the CP-nets without having to explicitly compute the induced partial orders. This new distance is defined as the Kendal Tau distance of the two $LexO$ linearizations of the partial orders.

Definition 2. *Given two O-legal CP-nets A and B, with m features, we define:*

$$CPD(A,B) = KT(LexO(A), LexO(B)) \tag{2}$$

We show that CPD is a distance over O-legal CP-nets.

Theorem 1. *Function CPD(A,B) satisfies the following properties:*

1. $CPD(A,B) \geq 0$*;*
2. $CPD(A,B) = CPD(B,A)$*;*
3. $CPD(A,B) \leq CPD(A,C) + CPD(C,B)$*.*
4. $CPD(A,B) = 0$ *if and only if* $A = B$*;*

Proof. Properties 1-3 are directly derived from the fact that KTD is a distance function over total orders. Let us now focus on property 4. In our context, $A = B$ if and only if they induce the same partial order. It is, thus, obvious that if $A = B$ then $CPD(A,B) = 0$ since $LexO(A) = LexO(B)$. Let us now assume that $A \neq B$. Thus A and B induce different partial orders. In principle, what could happen is that one partial order is a subset of the other. In such a case they would have the same $LexO$ linearizations and it would be the case that $CPD(A,B) = 0$, despite them being different. We need to show that this cannot be the case if A and B are O-legal. Let us first assume that A and B have the same dependency graph but that they differ in at least one ordering in one CP-table. It is easy to see that in such a case there is at least one pair of outcomes that are ordered in the opposite way in the two induced partial orders. Assume that A and B have a different dependency graph. Due to O-legality it must be that there is a least an edge which is present, say, in A and missing B. In this case by adding a non-redundant dependency we are reversing the order of at least two outcomes. □

We will now show how $CPD(A,B)$ can be directly computed from CP-nets A and B, without having to compute the linearizations. The computation comprises of two steps. The first step, which we call, normalization, modifies A and B so that each feature will have the same set of parents in both CP-nets. This means that each feature will have in both normalized CP-nets a CP-table with exactly the same number of rows corresponding each to the same assignment to its parents. The second step, broadly speaking, computes the contribution to the distance of each difference in the CP-table entries. We describe each step in turn.

Step 1: Normalization. Consider two CP-nets, A and B over m variables $V = \{X_1, \ldots, X_m\}$ each with binary domains. We assume the two CP-nets are O-legal with respect to a total order $O = X_1 < X_2 < \cdots < X_{m-1} < X_m$. We note that O-legality implies that the X_i can only depend on a subset of $\{X_1, \ldots, X_{i-1}\}$

Each variable X_i has a set of parents $Pa_A(X_i)$ (resp. $Pa_B(X_i)$) in A (resp. in B), and is annotated with a conditional preference table in each CP-net, denoted $CPT_A(X_i)$ and $CPT_B(X_i)$.

We note that, in general we will have that $Pa_A(X_i) \neq Pa_B(X_i)$. However, it is easy to extend the two CP-nets so that in both X_i will have the same set of parents $Pa_A(X_i) \cup Pa_B(X_i)$. This is done by adding redundant information to the CP-tables, which does not alter the induced ordering.

For example, let us consider $CPT_A(X_i)$, then we will add $2^{Pa_A(X_i) \cup Pa_B(X_i)} - 2^{Pa_A(X_i)}$ copies of each original row to $CPT_A(X_i)$, that is, one for each assignment to the variables on which X_i depends in B but not in A. After this process is applied to all the features in both CP-nets, each feature will have the same parents in both CP-nets and its CP-tables will have the same number of rows in both CP-nets. We denote with A' and B' the resulting CP-nets.

We note that normalization can be seen as the reverse process of CP-net reduction (Apt *et al.* 2008) which eliminates redundant dependencies in a CP-net.

Step 2: Distance Calculation Given two normalized CP-nets A and B, let $diff(A,B)$ represent the set of CP-table entries of B which are different in A and let $var(i) = j$ if CP-table entry i refers to variable X_j. Moreover, let $m = |V|$ and $flw(X)$ denote the number of features following X in order O. Let us define the two following quantities:

$$nSwap(A,B) = \sum_{j \in diff(A,B)} 2^{flw(var(j))+(m-1)-|Pa_B(var(j))|} \tag{3}$$

which counts the number of inversions that are caused by each different table entry and sums them up.

Theorem 2. *Given two normalized CP-nets A and B, we have:*

$$CPD(A,B) = nSwap(A,B) \tag{4}$$

We provide an example of how a difference in a CP-table entry affects the $LexO$ linearization.

Example 0.1. *Consider a CP-net with three binary features, A, B, and C, with domains containing f and $\overline{f}$ if F is the name of the feature, and with the cp-statements as follows: $a \succ \overline{a}$, $b \succ \overline{b}$, $c \succ \overline{c}$. A linearization of the partial order induced by this CP-net can be obtained by imposing an order over the variables, say Let variable ordering $O = A \succ B \succ C$. The $LexO(A)$ is as follows:*

$$\overbrace{\overbrace{abc \succ ab\overline{c}}^{B1Zone} \succ \overbrace{a\overline{b}c \succ a\overline{b}\overline{c}}^{B2Zone}}^{A1Zone} \succ \overbrace{\overbrace{\overline{a}bc \succ \overline{a}b\overline{c}}^{B3zone} \succ \overbrace{\overline{a}\overline{b}c \succ \overline{a}\overline{b}\overline{c}}^{B4zone}}^{A2zone}$$

Now, consider changing only the cp-statement regarding A to $\overline{a} \succ a$. Then, the linearization of this new CP-net can be obtained by the previous one by swapping the first outcome in the $A1zone$ with the first outcome in the $A2zone$, the second outcome in the $A1zone$ with the second outcome in the $A2zone$ and so on. Moreover, the number of swaps is directly dependent on the number of variables that come after A in the total order.

From Theorem 2 we can see that $0 \leq CPD(A, B) \leq 2^{m-1}(2^m - 1)$, where m is the number of features. In particular:

- $CPD(A, B) = 0$ when the two CP-nets have the same dependency graph and cp-tables and so they are representing the same preferences;
- $CPD(A, B) = 2^{m-1}(2^m - 1)$ when the two CP-nets have the same dependency graph but cp-tables with reversed entries, so they are representing preferences that are opposite to each other.

Notice that variables with different cp-statements in the representation give more value to the distance if they come first in the total order: the value decreases as the position in the total order increases. For instance it is easy to prove that if the cp-statement of the first variable in the total order differs, than $CPD \geq 2^{m-2}(2^m - 1)$.

Supporting Ethical Decisions

Ethical principles are modelled via a CP-net, say S, and an individual models her preferences via another CP-net, say B. We assume that these two CP-nets have the same features.

Of course this is a restriction and in general we think the features of these two CP-nets can overlap but not necessarily be the same. We are studying what happens when the two sets of features do not coincide. But for the purpose of this paper we will assume they do coincide.

Given the ethical principles and the individual's preferences, we need to guide the individual in making decisions that are not too unethical. To do this, we propose to proceed as follows:

1. We set two distance thresholds: one between CP-nets (ranging between 0 and 1), and another one between decisions (ranging between 1 and n).
2. We check if the two CP-nets A and B are less distant than t_1. In this step, we use CPD to compute the distance.
3. If so, the individual is allowed to choose the top outcome of his preference CP-net.
4. If not, then the individual needs to move down its preference ordering to less preferred decisions, until he finds one that is closer than t_2 to the optimal ethical decision. This is a compromise decision between what the preferences say and what the ethical principles recommend.

Empirical Analysis

We divide the empirical evaluation in two parts. Firstly, we evaluate the performances of the CPD distance by checking running time and deviation from the exact KTD distance. The first part of the experiments shows that in terms of computation time and error rate, our approximation performs extremely well. The second part of our experiments focuses on the ethical perspective. We show how the distance can be used in an ethical scenario to evaluate how much an individual decision maker deviates from an adopted ethical principle modeled as a CP-net.

Ethical Scenario

Given an ethical principle and the preference of an individual, both encoded as CP-nets, we want to understand if following the preferences will lead to an ethical action. Since in this scenario individuals want to act ethically, firstly the individual determines whether she can use her most preferred choice by checking if her CP-net is "sufficiently close" to the ethical CP-net. If these two CP-nets are farther apart than some threshold t_1, then we proceed down the preference ordering till we find a decision that is sufficiently close to the optimal ethical decision, according to another threshold t_2.

We represent the ethical principles with a CP-net A and the individual's preferences with a CP-net B, and we assume that these two CP-nets have the same features. We judge that the individual is acting ethically if $CPD(A, B) \leq t_1$. If yes, the individual knows that her preferences are pretty ethical and she can choose the best outcome induced by her CP-net.

If instead $CPD(A, B) > t_1$, we compute how many worsening flips we need to apply to her best decision (according to her preferences) to get to a decision that is closer than t_2 flips from the optimal ethical decision.

This empirical analysis is run varying n, t_1 and t_2, where n is the number of features, and t_1 and t_2 are the tolerances. We run experiments varying the number of features $2 \leq n \leq 8$. For each value of n we vary $t_1 \in \{0, 0.1, 0.2, 0.4, 0.8\}$. Low values of t_1 represents scenarios where the tolerance is absent or low. This means that, in order for a decision maker to take their first choice, they should have preferences very close to the ethical principle. Larger values of t_1 model less strict ethics, where people have more freedom of choice. For each value of n and t_1, we vary the value of t_2 ($2 \leq t_2 \leq (n + 2)/2$). This again represents scenarios where the freedom of individuals vary.

Given the values of n, t_1,, and t_2 we generate 1000 pairs of CP-nets (A, B) from a uniform distribution using the software described by (Allen *et al.* 2017; 2016). We compared values of the approximate CPD distance with the real KTD distance. This shows us how many times CPD is wrong and how much individuals need to sacrifice of their preferences in order to be ethical. We consider and report the following cases which represent the *confusion matrix* of our experiment:

1. True Positive (TP): $CPD(A, B) \leq t_1$ and $KTD(A, B) \leq t_1$. In this case, individual preferences are close to the ethical principles and decision makers choose their best alternative;
2. True Negative (TN): $CPD(A, B) > t_1$ and $KTD(A, B) > t_1$. In this case, individual preferences are not close to the ethical principles and the

decision makers must find a compromise;

3. False Positive (FP): $CPD(A, B) \leq t_1$ and $KTD(A, B) > t_1$. In this case, erroneously, individuals think they are acting ethically and consequently choose their best alternative even though it is not ethical;
4. False Negative (FN): $CPD(A, B) > t_1$ and $KTD(A, B) \leq t_1$. In this case, erroneously individuals think they are not acting ethically and they select a compromise decision even though they could select their top preferred decision.

The number of $TP + TN$ gives an idea of the accuracy of the distance; the higher this value, the higher confidence individuals can have in using the approximation of the distance to understand whether they are ethical or not.

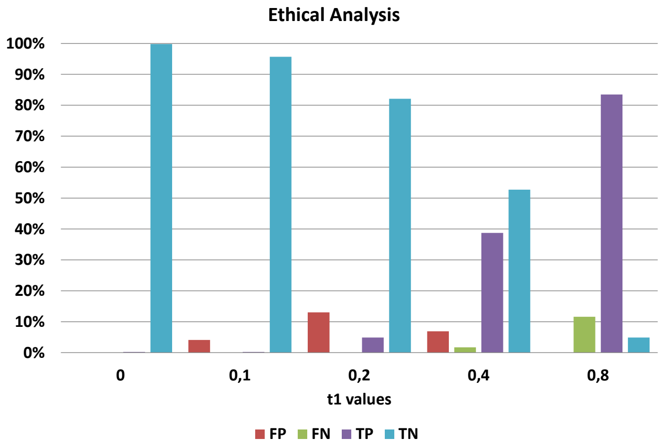

Figure 1: Percentage of TP, TN, FP, FN: the chart reports the number of cases for which CPD and KTD agree, or not, on the comparison based on the tolerance t_1. This gives an idea of the accuracy of the approximated distance.

Figure 1 shows the confusion matrix for $n = 7$ and $t_2 = 4$ while varying t_1. Notice that, as expected, when the tolerance t_1 is null or low, e.g., $t_1 = 0$ or $t_1 = 0.2$, individuals can almost never select their their best choice. Indeed, for $t_1 = 0$ the percentage of True Positives (purple bar) is close to 0% while for $t_2 = 0.2$ the percentage of True Positive is around 5%. This means that the decision makers preferences must be close to the ethical principle in order to have the freedom to choosing their best choice. Instead, when the tolerance is higher, they have more freedom to choose what they like. For example, with $t_1 = 0.4$, the percentage of True Positives (purple bar) is close to 40% while for $t_1 = 0.8$ it is more than 80%.

The next important question is: What happens when individuals cannot choose their first choice and have to look for another one which is closer to the ethical principles? Figure 2 reports the percentage of cases in which individuals have to find a compromise because their preferences are not close to the ethical principles, according to t_1,. For these cases we quantify the amount of compromise in terms of positions in the induced partial order. As before, when the tolerance is strict, an individual has to look for a compromise nearly every time. It is interesting to notice that the amount of compromise varies based on the value of t_2 and seems to be not influenced by t_1. This is quite natural, when $t_2 = 4$ it means that the individual has to find a choice that is in the top five positions of the ethical ordering in order to reach a compromise. This means that such a choice, on average, is in the first two positions of the individual's preference (red line in figure). The lower the value of t_2, the harder it becomes for the individual to find an ethical decision, and she has to descend down her preference order, on average, up to the fourth position to find an acceptable alternative.

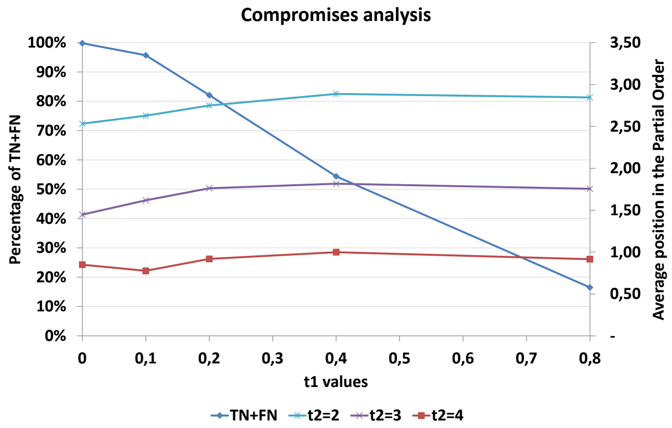

Figure 2: Compromises analysis: the charts reports a comparison between the number of times that individuals have preferences which are not close to the ethics and for which they have to look for a compromise and the quality of the compromise in terms of distance from their best choice.

Conclusions

In order to model and reason with both preferences and ethical principles in a decision making scenario, we have proposed a notion of distance between CP-nets, providing both a theoretical study and an experimental evaluation of its properties. We show that our approximation is both accurate in practice and efficient to compute.

Several extensions to our setting can be considered for the future. Indeed, we have made some assumptions on the two CP-nets for which we can compute the distance, that would be useful to relax. First, the two CP-nets over which we define the CPD distance have the same features, and with the same domains, but can differ in their dependency structure and CP-tables. It is important to also cover the case of CP-nets that may have different features and domains. Moreover, we have also assumed the two CP-nets are O-legal, that is, there is a total order of the CP-nets features that is compatible with the dependency links of both CP-nets. Intuitively, this means that the preferences are the ethical principles are not indicating completely opposite priorities. However, there could be situations where this is actually the case, and it is important to know how to combine preferences and ethical principles also in this case.

References

T.E. Allen, M. Chen, J. Goldsmith, N. Mattei, A. Popova, M. Regenwetter, F. Rossi, and C. Zwilling. Beyond theory and data in preference modeling: Bringing humans into the loop. In *Proceedings of the 4th International Conference on Algorithmic Decision Theory (ADT)*, 2015.

T.E. Allen, J. Goldsmith, H.E. Justice, N. Mattei, and K. Raines. Generating CP-nets uniformly at random. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

T. E. Allen, J. Goldsmith, H. E. Justice, N. Mattei, and K. Raines. Uniform random generation and dominance testing for cp-nets. *Journal of Artificial Intelligence Research*, 59:771–813, 2017.

Krzysztof R. Apt, Francesca Rossi, and Kristen Brent Venable. Comparing the notions of optimality in cp-nets, strategic games and soft constraints. *Ann. Math. Artif. Intell.*, 52(1):25–54, 2008.

Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

Craig Boutilier, Ronen Brafman, Carmel Domshlak, Holger Hoos, and David Poole. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2004.

Ronen I. Brafman and Yannis Dimopoulos. Extended semantics and optimization algorithms for CP-networks. *Computational Intelligence*, 20(2):218–245, 2004.

Y. Chevaleyre, U. Endriss, J. Lang, and N. Maudet. Preference handling in combinatorial domains: From AI to social choice. *AI Magazine*, 29(4):37–46, 2008.

Yann Chevaleyre, Frédéric Koriche, Jérôme Lang, Jérôme Mengin, and Bruno Zanuttini. Learning ordinal preferences on multiattribute domains: The case of CP-nets. In *Preference Learning*, pages 273–296. Springer, 2011.

Vincent Conitzer, Jérôme Lang, and Lirong Xia. Hypercubewise preference aggregation in multi-issue domains. In *22nd*, pages 158–163, 2011.

David Copp. *The Oxford Handbook of Ethical Theory*. Oxford University Press, 2005.

C. Cornelio, J. Goldsmith, N. Mattei, F. Rossi, and K.B. Venable. Updates and uncertainty in CP-nets. In *Proceedings of the 26th Australasian Joint Conference on Artificial Intelligence (AUSAI)*, 2013.

C. Cornelio, U. Grandi, J. Goldsmith, N. Mattei, F. Rossi, and K.B. Venable. Reasoning with PCP-nets in a multi-agent context. In *Proceedings of the 14th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2015.

C. Domshlak, E. Hüllermeier, S. Kaci, and H. Prade. Preferences in AI: An overview. 175(7):1037–1052, 2011.

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. Comparing partial rankings. *SIAM J. Discret. Math.*, 20(3):628–648, March 2006.

J. Goldsmith, J. Lang, M. Truszczyński, and N. Wilson. The computational complexity of dominance and consistency in CP-nets. *Journal of Artificial Intelligence Research*, 33(1):403–432, 2008.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Jerome Lang and Lirong Xia. Sequential composition of voting rules in multi-issue domains. *Mathematical Social Sciences*, 57(3):304–324, 2009.

A. Loreggia, N. Mattei, F. Rossi, and K.B. Venable. On the distance between CP-nets. In *Proceedings of the 17th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2018.

N. Mattei and T. Walsh. PrefLib: A library for preferences, HTTP://WWW.PREFLIB.ORG. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT)*, 2013.

N. Mattei, M. S. Pini, F. Rossi, and K. B. Venable. Bribery in voting with CP-nets. *Annals of Mathematics and Artificial Intelligence*, 68(1–3):135–160, 2013.

G. Pigozzi, A. Tsoukiàs, and P. Viappiani. Preferences in artificial intelligence. 77:361–401, 2015.

F. Rossi, K.B. Venable, and T. Walsh. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Morgan and Claypool, 2011.

Ganesh Ram Santhanam, Samik Basu, and Vasant Honavar. Verifying preferential equivalence and subsumption via model checking. In Patrice Perny, Marc Pirlot, and Alexis Tsoukiàs, editors, *Algorithmic Decision Theory - Third International Conference, ADT 2013, Bruxelles, Belgium, November 12-14, 2013, Proceedings*, volume 8176 of *Lecture Notes in Computer Science*, pages 324–335. Springer, 2013.

Amartya Sen. *Choice, Ordering and Morality*. Blackwell, Oxford, 1974.

Judith Jarvis Thomson. The trolley problem. *The Yale Law Journal*, 94(6):1395–1415, 1985.