18 Value Alignment via Tractable Preference Distance

Andrea Loreggia, Nicholas Mattei,
Francesca Rossi, and K. Brent Venable

CONTENTS

INTRODUCTION

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. Hence, the study of preferences in computer science and AI has been very active for a number of years with important theoretical and practical results [13,21] as well as libraries and datasets [20]. In many scenarios including multi-agent systems [25] and recommender systems [22], user preference play a key role in driving the decisions the system makes. Thus it is important to have preference modeling frameworks that allow for expressive and compact representations, effective elicitation techniques, and efficient reasoning and aggregation.

If we want people to trust AI systems, we need to provide these systems with the ability to discriminate between what one would broadly call "good" and "bad" decisions. In many instances, the quality of a decision should not be based only on the preferences or optimization criteria of the decision makers, but also on properties related to the impact of the decision such as whether or not it is ethical or legal according to constraints or priorities given by any number of exogenous sources [5,24,26]. Indeed, there may be specific ethical principles, depending on the context, that could and should override the subjective preferences of the decision maker.

For the subjective preferences, they may apply to one or more of the individual components of a complex decision, rather than to the whole thing. For example, if we need to choose a car, we may prefer certain colors 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 [6] is a convenient and expressive way to model preferences that has been used widely in the preference handling community [8,12,15,23]. CP-nets provide an effective, compact

way to qualitatively model preferences over decisions (often called outcomes) with a combinatorial structure. CP-nets are also easy to elicit and provide efficient methods for optimization, search and reasoning [2,9]. Moreover, in a collective decision-making scenario, several CP-nets can be aggregated, for example, using voting rules [10,19], 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 are not the only source of information we should consider [5,24,26]. Indeed, depending on the context, we may have to consider specific ethical principles or laws derived from an appropriate ethical theory or local statutes [11]. While preferences are important, when preferences and ethical principles (constraints) are in conflict, these constraints 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, for example, ensuring gender and minority diversity. Therefore, it is essential to have systematic and rigorous methodologies to evaluate if preferences are compatible with a set of ethical principles, and to measure the difference between the preferences and 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 chapter we argue that we can use CP-nets to model both exogenous priorities, for example, those provided by ethical principles, as well as the subjective preferences of decision makers. Thus, the distance between an individual's subjective preferences and the exogenous ethical principles can be measured via a notion of distance between CP-nets. Indeed, we use a notion of distance (formally a distance function or metric) between CP-nets first discussed by [18] to evaluate the decisions made by autonomous systems. This metric, called KTD, generalizes the classic [16] τ (KT) distance often used to measure the distance between partial orders. KTD works by counting the number of inverted pairs between two complete, strict linear orders while adding a penalty parameter p defined for partial rankings [14].

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 in general computing a distance between these induced partial orders is computationally intractable. Therefore, for a practical AI system we propose using a tractable approximation, called O-CPD by [18], that is computed directly over the CP-nets dependency graphs. In this chapter we study the quality of the approximation for decision making and describe a value alignment procedure that uses O-CPD.

The value alignment procedure we propose computes the distance between subjective preferences and ethical principles, and makes decisions using the subjective preferences only if they are *close enough* to the ethical principles, where being *close enough* depends on a threshold over CP-net distances. If instead the preferences diverge too much from the ethical principles, we move to a less preferred decision until we find one that is a*satisfactory compromise* between the ethical principles and the user preferences. The compromise is defined by setting a second threshold over distances between decisions.

As mentioned above, O-CPD is an approximation of KTD, thus in general the values returned by these two distances may be different. In Reference 18 we give some theoretical bounds for the error, but they are fairly wide and may not work well in practice as a value alignment procedure. Hence, here we perform an experimental evaluation showing that the approximation almost always agrees with the real value, in terms of positioning the distance compared to the set threshold, and that the quality of the decision with respect to the subjective preferences does not significantly degrade, that is, 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, short for Conditional Preference networks, were first proposed by [6] are a graphical model for compactly representing conditional and qualitative preference relations. CP-nets are comprised of 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* to the the parent variables, $Pa(x)$. 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 $Pa(x) = \{x_1, \ldots, x_n\}$, $D(x) = \{a_1, \ldots, a_m\}$, for each $x_i \in Pa(x)$, $x_i = v_i$ is an assignement to a parent of x, with $v_i \in \mathcal{D}(x_i)$ 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, depicted graphically in Figure 18.1, 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 [6]. 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

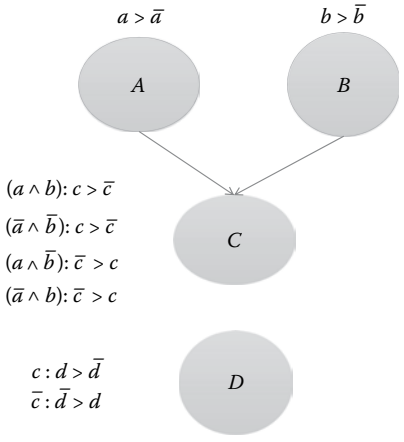

FIGURE 18.1 CP-net with four features A, B, C, and D, with binary domains. Next to each variable there is the correspondent CP-table.

* Notice that CP-nets is written with capital CP, while cp-statements is written with small cp, since they mean different things and since the literature introducing such notions uses this notation.

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 [7]: 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 work we want to compare CP-nets while leveraging the compactness of the representation. To do this, we consider a profile (P, O), where P is a collection of n CP-nets whose graph is a directed acyclic graph (DAG) over m common variables with binary domains and O is a total order over these variables. We require that the profile is O-legal [17], 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 the ordering O.

Since every acyclic CP-net is satisfiable [6], 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 by [6] 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 has 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

```
1:  function Lin(A, O, LexO, o) ▷ Where A is a CP-net, O is the O-legal order on A, LexO: is
    the linearization we compute (empty to start with), and o is an outcome.
2:  if O = Null then
3:     LexO.append(o)
4:     return LexO
5:  end if
6:  v = pop(O)
7:  for value ∈ CPT_{A,Par(v)}(v) do
8:     temp = o + value
9:     LexO = Lin(A, O, LexO, temp)
10:  end for
11:  return LexO
12: end function
```

USING CP-NETS TO MODEL ETHICS

In this section, we describe through an example how CP-nets can be used to model ethical principles that may come from one or more ethical theories or societal value systems [11]. We start by modeling a possible scenario where autonomous vehicles operate in environments where both artificial agents

and humans coexist [5]. In such a scenario, each human driver can have his/her own subjective preferences or priorities over actions to take in specific situations for which the traffic laws do not prescribe a specific behavior. Moreover, there can be collective ethical guidelines that a community may come up with, and would like all drivers to follow, with some tolerance. Many examples of situations like this, where personal preferences and ethical preferences collide, are collected and used as dilemmas, for example at the Moral Machines website [4]. Most of these dilemmas are derived from the classic Trolley roblem [26], which we use as inspiration in our examples.

We propose a value alignment procedure which allows the AI system to act in compliance with the societally imposed ethical principles. We discuss how the proposed value alignment procedure, based on CP-nets, assures that inconsistent behaviors would be prevented by autonomous vehicles. Suppose the vehicle has a brake malfunction while approaching an intersection where pedestrians of both a dog and human variety are crossing the street. The driver has two options: go straight ahead or swerve. If the driver continues to go straight, she then has the unfortunate option of running over either the group of dogs or the group of humans, roughly choosing how many of them will be injured but resulting in saving all of the passengers in the vehicle. On the other hand, the driver can swerve off the road, which will result in saving both the dogs and the human pedestrians, but injuring all of the passengers.

Figure 18.2 shows the preference of a diligent driver (called the Angel Driver) as both a CP-net and the resulting induced ordering over the possible actions. If this driver would find herself in the situation we have just described, her preferences show that deciding to swerve, and thus injuring all of her passengers, depends on the number and type of pedestrians in the intersection. Indeed, she prefers to go straight, resulting in running over a small group of dogs and saving both her passengers and the human pedestrians. Hence, the decision of the driver is primarily influenced by the type of pedestrians and how many of them will be injured.

These preferences, modeled in the CP-net on the left side of Figure 18.2, induce an ordering in the space of all the possible actions that is depicted on the right side. This ordering goes from the

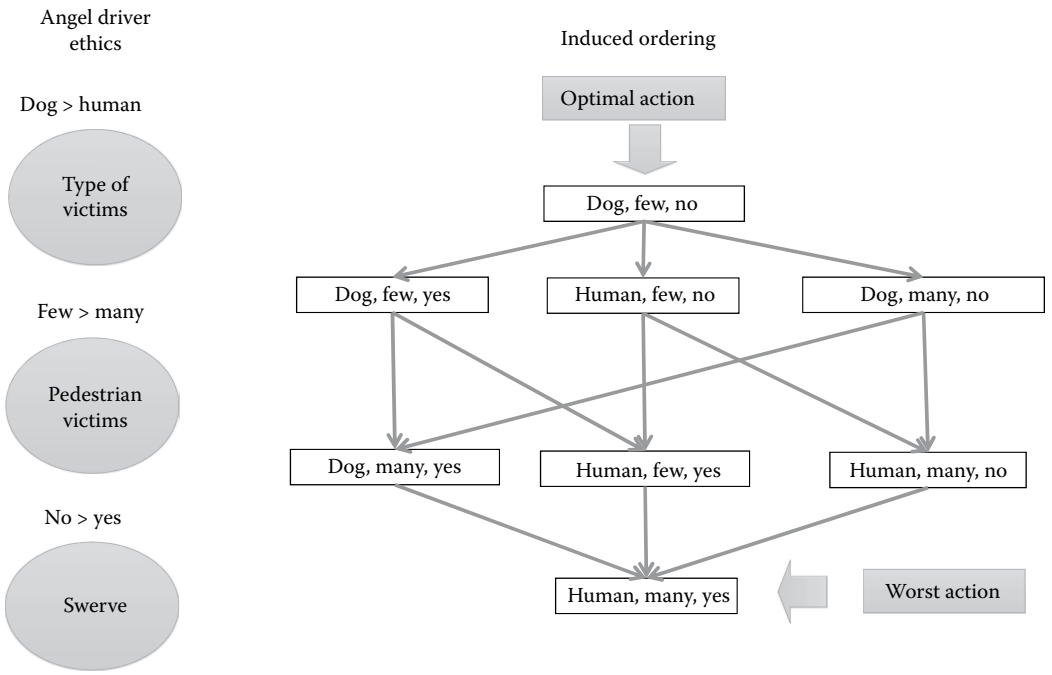

FIGURE 18.2 On the left side: a CP-net that models the moral preferences of a diligent driver. On the right side: the induced partial order over the action space.

optimal (most preferred) action (i.e., the driver prefers not to swerve, running over a small group of dogs and saving her passengers), to the worst (least preferred) action (i.e., running over a large group of humans and swerving, which injures all of the passengers).

In Figure 18.3, we show the preference of a different kind of driver, a *Devil* driver. Devil drivers are driven by hostility and consequently prefer to run over the greatest number of human pedestrians. As the reader can notice, relationships between outcomes (which are represented by the arrows) drastically change, resulting in different optimal and worst actions according to the preferences of the Devil driver.

Yet another set of preferences over this space may come from what the relevant community, or the society, says should be the ethical principles. Figure 18.4 compactly describes moral preferences over the possible solutions which we assume are derived from some appropriate ethical theory or have been decided upon through a collective effort in a society [4,11]. For our example, we assume the collectivity prefers not to kill human pedestrians and save as many lives as possible. In cases where killing someone is unavoidable, then we assume it is morally preferable to have the smallest number of victims. Finally, we assume it is acceptable to possibly cause injuries on the passengers when there is a sufficiently large number of human pedestrians in danger in the street.

Suppose that users of an autonomous vehicle can interact with the system by specifying their preferences over scenario that the vehicle may encounter. Preferences of an Angel driver and a Devil driver induce partial orders which are very different and can lead to different behaviors if specified to the vehicle. Specifically, when faced with our example scenario, an Angel driver would always prefer to run over a small group of dogs, saving most of the dogs and all of the human pedestrians and passengers. As we move down the preference ordering of the Angel driver through a sequence of worsening flips, the actions we find result in more serious consequences as we move toward the worst outcome. On the other hand, a Devil driver has the opposite ordering over the possible actions. Facing the same scenario the Devil driver would prefer to run over the largest group of humans.

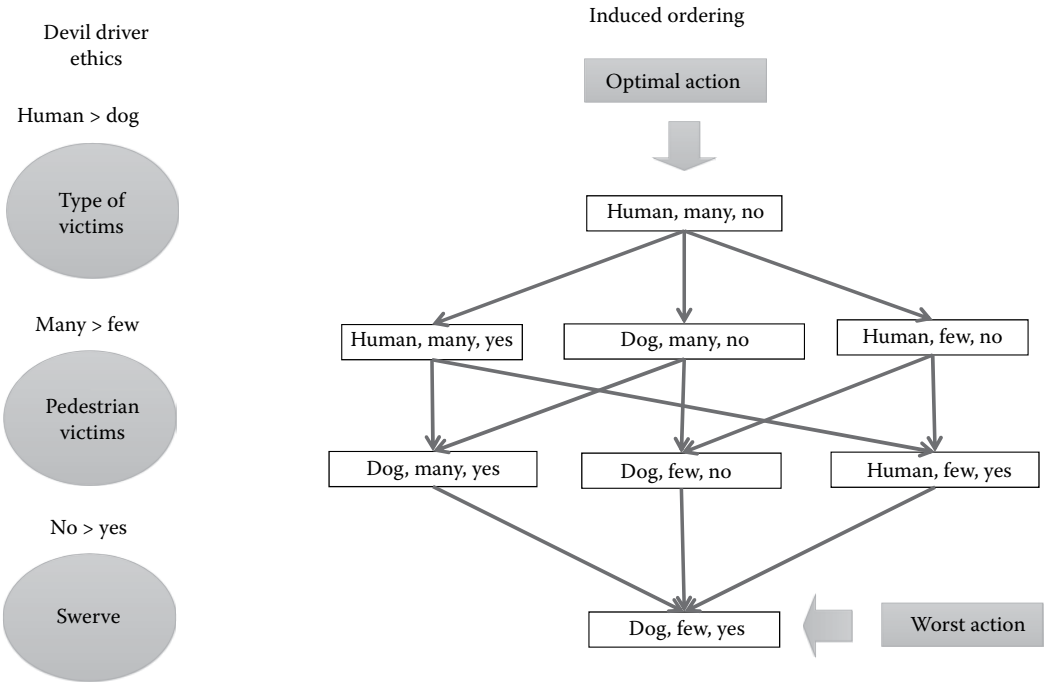

FIGURE 18.3 On the left side: a CP-net that models the moral preferences of a Devil driver. On the right side: the induced partial order over the action space.

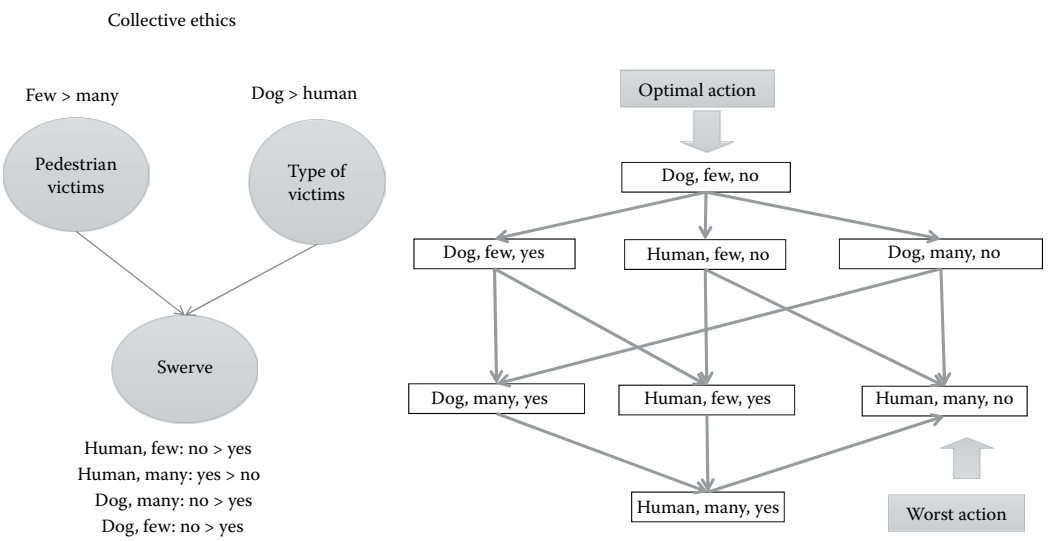

FIGURE 18.4 On the left side: a CP-net that models the collective moral preferences of a society. On the right side: the induced partial order over the action space. Note that we have kept the outcomes in the same positions as Figure 18.2 and rearranged the arrows which show the ordering according to worsening flips.

Indeed, he also prefers solutions that are maximally damaging and could be even more serious, for instance, by swerving the car and also injuring his passengers.

Hence it is important to have tools which are able to understand how different or how similar two given preference orderings are. This is needed whether the preferences represent moral principles, exogenous priorities, or individual preferences. Having such metrics allows us to define value alignment procedures that enable artificial agents to both follow established guidelines and correct or prevent harm from non-compliant behaviors. When preferences and priorities are modeled as CP-nets, a straightforward metric that can be used to measure the distance between two preferences is the [16] τ (KT) distance, which counts how many swaps are required in order to change one linear order into another one. [18] proposed such a metric for CP-nets, along with tractable approximations. In the next section, we discuss these metrics and define a value alignment procedure for supporting decision making under exogenous priorities.

NOTIONS OF DISTANCE BETWEEN CP-NETS

CP-nets do not induce total orders, but rather a partial order over the solution space, meaning that some outcomes can be incomparable as seen in the examples in the last section. The following definition is an extension of the [16] τ (KT) distance with a penalty parameter p defined for partial rankings by [14]. A more complete treatment and proofs of correctness for this definition are discussed by [18].

Definition 1 (Kendall τ Distance (KTD)): Given two CP-nets A and B inducing partial orders P and Q over the same set of outcomes U:

$$KTD\ (A,B) = KT(P,Q) = \sum_{\forall i,j \in U, i \neq j} K^{p}_{i,j}(P,Q) \tag{18.1}$$

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).

This metric gives a formal measure of a distance between two partial orders. Since this is a straightforward extension of the popular KT distance, which we will call KTD, we use it as a basis for comparisons. Unfortunately, this measure is computationally difficult to compute [18]. Since the domains where we would like to apply this distance are combinatorial, the running time required to compute the KTD can be exponential and hence, intractable. For this reason, it is desirable to use approximations which are more efficient in terms of processing time. However, using approximations may introduce some error and we would like these errors to be as small as possible.

A novel approximation of KTD is defined by [18] and we will use this definition here to study how well a system can support decisions using an approximate distance. For the sake of readability, we report only the definition of the approximation metric, called O-CPD, below. We refer the reader to the original work for details and proofs of bounds and correctness [18].

Definition 2 (O-Legal CP-net Distance (O-CPD)): Given a variable X_i, the function $flw(X_i)$ returns the number of variables following X_i in the ordering O. We thus define O-CPD as

$$\text{O-}CPD\ (A,B) = \sum_{j \in diff(A,B)} 2^{flw(var(j))+(m-1)-|Pa_B(var(j))|}. \tag{18.2}$$

The O-CPD metric is ideal as it has both an exponentially smaller computation time (on average) and provides provable error bounds with respect to KTD. In addition, O-CPD leverages the compact representation provided by CP-nets and is therefore more memory efficient to compute. For all these reasons, we think it is an ideal distance measure that is useable for general AI systems. In our empirical experiments, we will study how different O-CPD and KTD can be, and whether or not O-CPD can be used to judge the value alignment of CP-nets.

USING CP-NETS TO SUPPORT ETHICAL DECISIONS

Suppose that ethical principles are modeled via a CP-net N_e and an individual (human or computer) agent models her preferences via another CP-net N_p where both N_e and N_p have the same set of 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. An interesting direction for future work is to expand the notions of distance and value alignment to cases where the two sets of features do not coincide. However, for the purpose of this paper we will assume they are the same.

Given the ethical principles and the agent preferences, we need to guide the agent in making decisions that are morally acceptable according to the given ethical principles described by N_e. To do this, we propose the following value alignment procedure:

1. We set two distance thresholds: one between CP-nets, called t_1, that ranges between 0 and 1, and another between actions, called t_2, that ranges between 1 and n (the number of features of both N_e and N_p).
2. We check if the distance between CP-nets N_e and N_p is less than t_1. Here, we use O-CPD to compute the distance in a tractable way.

3. If the distance is below the threshold t_1, the agent is allowed to choose the top action of his preference CP-net.
4. If the distance is above the threshold t_1, then the agent must move down his preference ordering, through worsening flips, to less a preferred action, until he finds an action that is closer than t_2 to the optimal action according to the ethical CP-net N_e. This is a compromise decision between what the preferences say and what the ethical principles recommend.

The fundamental idea in this value alignment procedure is that agents can behave as they prefer *only* when their preferences are close enough to the specified ethical principles. Otherwise, the agent must compromise by finding a solution that is closer to the one suggested by the ethics. Turning to our examples in section "Using CP-Nets to Model Ethics," an autonomous vehicle configured to use the preferences of an Angel driver would act as prescribed by their individual preferences since these preferences are close enough to the collective ethical principles. In the same scenario, however, an autonomous vehicle configured to use Devil driver preferences will prevent the maladaptive behavior of running over a large group of humans by searching for a more acceptable solution based on the tolerance t_2 that the vehicle is configured to use.

EMPIRICAL ANALYSIS

In this section, we show how we can use the CP-net distance metrics in an ethical scenario to evaluate how much an individual decision maker (agent) deviates from an adopted ethical principle modeled as a CP-net. Given an ethical principle and the preference of an agent, both encoded as CP-nets, we want to understand if following the preferences will lead to an ethical action as defined by the ethical CP-net. Since in this scenario agents want to act ethically, the agent first 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. Hence, t_1 and t_2 represent upper bounds about how far an agent may deviate from the ethical principles.

We represent the ethical principles with a CP-net N_e and the preferences of the agent with a CP-net N_p, and we assume that these two CP-nets have the same features. We say that the individual is acting ethically if $O\text{-CPD}(N_e, N_p) \leq t_1$. If indeed the agent is acting ethically, she can choose the best outcome induced by her CP-net. If instead $O\text{-CPD}(N_e, N_p) > t_1$, we compute how many worsening flips we need to apply to her best action (according to her preferences) to arrive at an action that is closer than t_2 flips from the optimal ethical decision.

We run an empirical analysis 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 represent 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 agents to do whatever they desire varies.

Given the values of n, t_1, and t_2, we generate 1000 pairs of CP-nets (N_e, N_p) from a uniform distribution using the software described by [1,3]. By comparing values of the approximated distance O-CPD with the real distance KTD, we can see both how many times O-CPD is wrong as well as how many individuals need to compromise on their preferences in order to be ethical. We consider the following cases, which represent the *confusion matrix* of our experiment:

1. True Positives (TP): $O\text{-CPD}(N_e, N_p) \leq t_1$ and $KTD\ (N_e, N_p) \leq t_1$. In this case, the agent preferences are close to the ethical principles and decision makers choose their best alternative.

2. True Negatives (TN): O-CPD(N_e, N_p) $> t_1$ and KTD (N_e, N_p) $> t_1$. In this case, the agent preferences are not close to the ethical principles and the decision makers must find a compromise.
3. False Positives (FP): O-CPD(N_e, N_p) $\leq t_1$ and KTD (N_e, N_p) $> t_1$. In this case, erroneously, the agent thinks they are acting ethically and consequently choose their best alternative even though it is not ethical.
4. False Negatives (FN): O-CPD(N_e, N_p) $> t_1$ and KTD (N_e, N_p) $\leq t_1$. In this case, erroneously, the agent thinks they are not acting ethically and they select a compromise decision even though they could select their top preferred decision.

The number of $TP + TN$ provides us with a measure of the accuracy of the respective distance metric; the higher this value, the higher the confidence individuals can have in using the approximation of the distance. This means that the higher is the value the higher is the likelihood they are behaving correctly with respect to the real distance between their preferences and the ethical principle. Looking at Figure 18.5 for $t_1 = 0$ gives us an experimental proof of the formal statement (proven in Reference 18) that O-CPD and KTD converge when the distance is 0, that is, we never make any errors when we must act exactly as the ethical constraints prescribe.

Figure 18.5 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 small, for example, $t_1 = 0$ or $t_1 = 0.2$, individuals can almost never select 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 of taking their best choice. Instead, when the tolerance is higher, the agent has more freedom to choose what he/she likes. 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 we consider is: What happens when agents cannot take their optimal decision and thus must look for another one closer to the ethical principles?

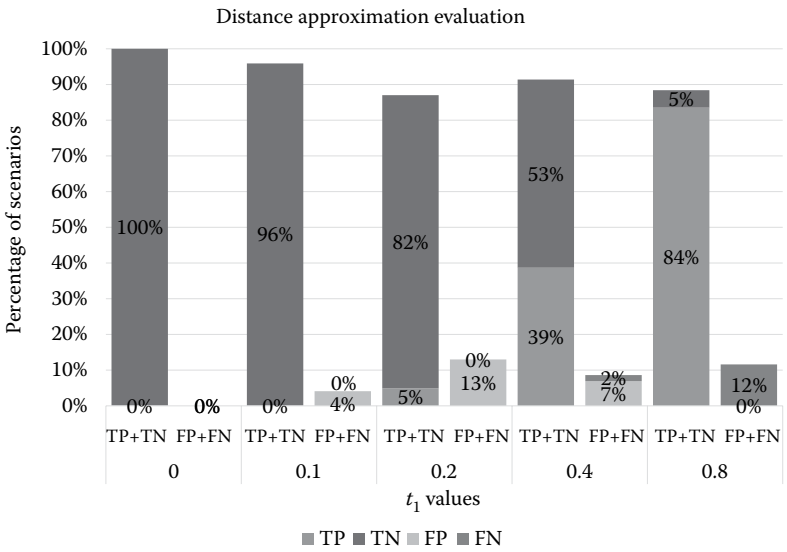

FIGURE 18.5 Percentage of TP, TN, FP, FN: different values of t_1 are reported on the x-axis. For each value the chart reports the number of cases for which O-CPD and KTD agree (TP or TN), or not (FP or FN), with respect to the value of the tolerance t_1. Note that O-CPD and KTD agree in over 80% of the cases across all values of t_1.

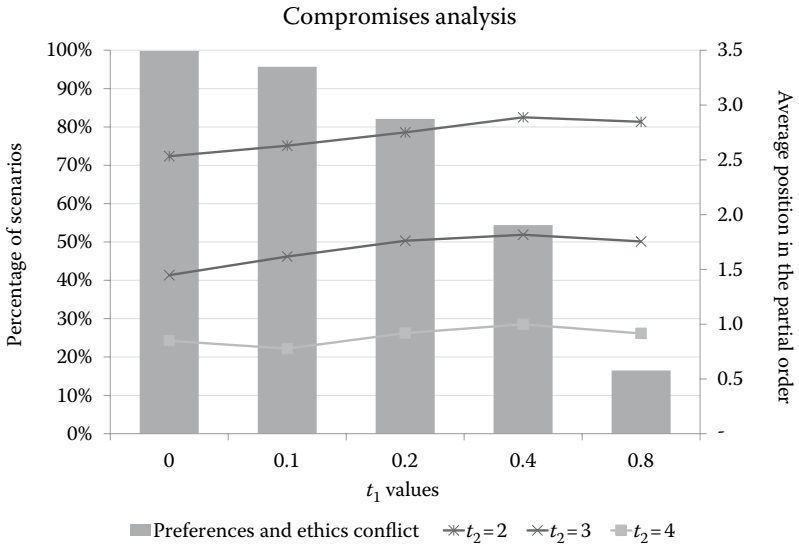

FIGURE 18.6 Compromise Analysis: in the chart x-axis reports different values of t_1, for each value of t_1, a bar shows the percentage of times an agent has to look for a compromise because her preferences are far from the ethical principle. Moreover, for each value of t_1 different lines, one for each tested value of t_2, shows the quality of the compromise as measured by the distance between the compromise and the best choice of the individual.

Figure 18.6 reports the percentage of cases in which agents have to find a compromise because their preferences are not close enough to the ethical principles, according to the threshold t_1. In the chart, orange bars report the percentage of times an agent has to look for a compromise for the specific value of t_1. For example, for $t_1 = 0.1$, it is about 95%.

For each of the considered values of t_1, we also tested the quality of the compromise in terms of distance between the compromise and the best choice of the individual. Different lines report the average position of the chosen outcome for different values of t_2. For these cases, we quantify the amount of compromise by counting the positions in the induced partial order. As before, when the tolerance is strict, an agent has to look for a compromise nearly every time. It is interesting to note that the amount of compromise varies based on the value of t_2 and seems to not be influenced by t_1. However, upon further reflection this result is quite natural, when $t_2 = 4$ it means that the agent 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 agent preferences (red line in Figure 18.6). The lower the value of t_2, the harder it becomes for the agent 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.

CONCLUSIONS

To create AI systems that either make autonomous decisions or support human decision makers, we must ensure that the systems themselves are aware of the ethical principles that are involved. In this chapter, we have argued that we can use CP-nets to model both the subjective preferences as well as the exogenous ethical and legal priorities that bear on a decision. Moreover, we have considered a notion of distance between CP-nets that is tractably computable and we have shown how to use it to define a value alignment procedure that checks if the preferences are *close enough* to the ethical principles. Finally, we have provided an experimental evaluation showing that the quality of the decisions, defined in relation to the subjective preferences) does not significantly degrade when conforming to the ethical principles.

Several interesting extensions to our setting can be considered to create more useable and capable AI systems. 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, in this paper, the two CP-nets have the same features, with the same domains, and can only differ in their dependency structure and CP-tables. It is important to extend our work to cover the case when CP-nets have different features and domains. Second, our work assumes that all preference orders in the CP-nets are strict. Relaxing this assumption to include impartiality, as well as incomparability, in the CP-statements is an important step to extending our work to more practical situations.

ACKNOWLEDGMENTS

This work is partially supported by the project "Safety Constraints and Ethical Principles in Collective Decision Making Systems" which is funded by the Future of Life Institute.

REFERENCES

1. 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.
2. 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)*, Lexington, KY, pp. 3–18, 2015.
3. 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)*, Phoenix, Arizona, pp. 872–878, 2016.
4. http://moralmachine.mit.edu/
5. J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.
6. C. Boutilier, R. Brafman, C. Domshlak, H. Hoos, and D. Poole. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2004.
7. R. I. Brafman, and Y. Dimopoulos. Extended semantics and optimization algorithms for CP-networks. *Computational Intelligence*, 20(2):218–245, 2004.
8. 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.
9. Y. Chevaleyre, F. Koriche, J. Lang, J. Mengin, and B. Zanuttini. Learning ordinal preferences on multiattribute domains: The case of CP-nets. In Johannes Fürnkranz Eyke Hüllermeier, editor, *Preference Learning*, Springer, Berlin Heidelberg, pp. 273–296, 2011.
10. V. Conitzer, J. Lang, and L. Xia. Hypercubewise preference aggregation in multi-issue domains. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, AAAI Press, Barcelona, Catalonia, pp. 158–163, 2011.
11. D. Copp, editor. *The Oxford Handbook of Ethical Theory*. Oxford: Oxford University Press, 2005.
12. 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)*, Istanbul, Turkey, pp. 969–977, 2015.
13. C. Domshlak, E. Hüllermeier, S. Kaci, and H. Prade. Preferences in AI: An overview. *Artificial Intelligence*, 175(7):1037–1052, 2011.
14. R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM Journal of Discrete Mathematical*, 20(3):628–648, March 2006.
15. 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.
16. M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
17. J. Lang and L. Xia. Sequential composition of voting rules in multi-issue domains. *Mathematical Social Sciences*, 57(3):304–324, 2009.

18. 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)*, Stockholm, Sweden, 2018.
19. N. Mattei, M. S. Pini, F. Rossi, and K. B. Venable. Bribery in voting over combinatorial domains is easy. In *Proceedings of the 11th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Valencia, Spain, pp. 1407–1408, 2012.
20. 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)*, Bruxelles, Belgium, pp. 259–270, 2013.
21. G. Pigozzi, A. Tsoukiàs, and P. Viappiani. Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 77:361–401, 2015.
22. F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.
23. F. Rossi, K. B. Venable, and T. Walsh. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Morgan and Claypool, 2011.
24. A. Sen. *Choice, Ordering and Morality*. Blackwell, Oxford, 1974.
25. Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.
26. J. J. Thomson. The trolley problem. *The Yale Law Journal*, 94(6):1395–1415, 1985.