The AAAI-17 Workshop on
AI, Ethics, and Society
WS-17-02

Modelling Ethical Theories Compactly

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

Francesca Rossi
IBM Research and University of Padova
frossi@it.ibm.com

K. Brent Venable
Dept. of Computer Science
Tulane University
kvenabl@tulane.edu

Abstract

Recently a large attention has been devoted to the ethical issues arising around the design and the implementation of artificial agents. This is due to the fact that humans and machines more and more often need to collaborate to decide on actions to take or decisions to make. Such decisions should be not only correct and optimal from the point of view of the overall goal to be reached, but should also agree to some form of moral values which are aligned to the human ones. Examples of such scenarios can be seen in autonomous vehicles, medical diagnosis support systems, and many other domains, where humans and artificial intelligent systems cooperate. One of the main issues arising in this context regards ways to model and reason with moral values. In this paper we discuss the possible use of AI compact preference models as a promising approach to model, reason, and embed moral values in decision support systems.

Introduction

Nowadays scenarios where humans and intelligent agents collaborate together to reach a common decision are growing in many different disciplines and real-life situations. For instance, the increasing number of autonomous vehicles driving around force us to think about the implications of the meaning of making autonomous decision. Such intelligent agents could face several situations where they have to resolve moral dilemmas, such as in the well know trolley problem for autonomous vehicles (see for example (Thomson 1985) for a description of several of these situations).

Artificial agents helping professionals shall agree to some form of deontological code for that profession. Think for instance to medical scenarios where humans and intelligent agents collaborate together to find a therapy. Doctors agree to the Hippocratic oath and would not trust suggestions coming from artificial agents that do not follow the same principles of the oath.

Agents should agree to the same ethical principles as humans in the same scenario. Thus it is crucial to be able to model ethical principles in a way that they can be effectively used by artificial agents. Researches from different areas have already studied different frameworks to model and to enable agents making decisions. Autonomous agents already schedule activities according to safety constraints, or to make a collective decision using some form of voting protocol which tries to satisfy the subjective preferences of all members of the decision makers group. We propose to adapt some of these frameworks to embed moral values in the decision process.

Compactness in modelling both preferences and moral values is a necessity when it comes to implementations for artificial agents. Humans are very good at abstracting away details which are not relevant for decision making and perceive as atomic even complex events or objects which would require large amounts of details to be formally described. Artificial agents don't have this luxury. They rely on combinatorial structures for the vast majority of the knowledge they acquire and store. This is true also when it comes to preferences and a key challenge that has been tackled by the area of knowledge representation has been that of mapping orderings over large sets of options into compact (graphical) models while trying to minimize the information that is lost in doing so.

Since ethical principles define the same kind of structures as preferences that is, priority orderings over the possible decisions, it is reasonable to conjecture that also ethical requirements will need to be modelled compactly in order to be embedded into a machine. One may argue that there are alternatives available. For example, one could take a machine learning based approach where "ethics" is modelled by one or more learning modules trained on, for example, dilemmas and corresponding solutions. While this approach may be feasible, it does raise some concerns. For example, it may not be acceptable that the artificial agent will not be able to provide an explanation on why it judged one action "more ethical" than another. Moreover, as noted in many papers in the literature, e.g. (Allen, Varner, and Zinser 2000), bottom up approaches to ethics tie the results to the data on which the module is trained. This may lead to undesirable outcomes if the data is biased or not general enough.

Background

We organize our discussion as follows. We consider the most popular compact preference models and, for each of them, after providing a short background we discuss issues and research challenges related to adopting them as model for modelling and reasoning with ethical theories.

Hard and Soft Constraints

Hard constraints, usually just called constraints, model restrictions on the combination of values that some decision variable can take. For example, in a scenario where we need to schedule activities over time, we may use one decision variable for each activities, which can take values from the time line, and we may pose the constraint that activity A has to occur before activity B. Thus, with a hard constraint, each combination of values of variables A and B is either feasible (if it satisfies the constraint) or not. Given several of such constraints, the global scenarios that are declared as feasible are those in which all constraints are satisfied.

Soft constraints generalise the notion of constraints to allow for more than just two states (feasible or not) for the value combinations. More precisely, a *soft constraint* (Rossi, van Beek, and Walsh 2006) involves a set of decision variables and associates a value from a (totally or partially ordered) set to each instantiation of its variables. Such a value is taken from a preference structure $\langle A, +, \times, 0, 1\rangle$, where A is the set of preference values, $+$ induces an ordering over A (where $a \leq b$ iff $a+b=b$), $\times$ is used to combine preference values, and 0 and 1 are the worst and best element. For example, in the activity scheduling example described above, we may work with a preference structure which includes totally ordered values from 0 to 1, where a higher value denotes a higher preference, and we may have a soft constraint assigning value 0 to combinations of values (A=a,B=b) where $a \not\leq b$, and value $(b-a)/b$ to the other combinations of values, meaning that we do not allow A to occur after B, and when A is before B, we prefer these two activities to be as close as possible.

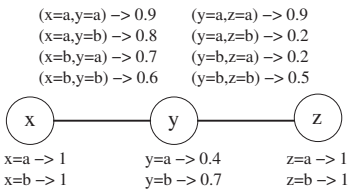
A *Soft Constraint Satisfaction Problem* (SCSP) is a tuple $\langle V, D, C, A\rangle$ where V is a set of variables, D is the domain of the variables, and C is a set of soft constraints (each one involving a subset of V) associating values from A.

An instance of the SCSP framework is obtained by choosing a specific preference structure. For instance, a *classical CSP* (Dechter 2003) is just an SCSP where the preference structure is $S_{CSP} = \langle\{false, true\}, \vee, \wedge, false, true\rangle$. Preference values are only true and false and they are combined via logical and.

Fuzzy CSPs (Rossi, van Beek, and Walsh 2006) are instead modeled choosing $S_{FCSP} = \langle[0,1], max, min, 0, 1\rangle$ that means that preference values are in $[0,1]$ and we want to maximize the minimum preference value. Fuzzy CSPs are useful when we have safety-critical applications, since we focus on the worst preference value when we evaluate a complete variable assignment.

For weighted CSPs, the c-semiring is $S_{WCSP} = \langle R^+, min, +, +\infty, 0\rangle$: preferences are interpreted as costs from 0 to $+\infty$, and we want to minimize the sum of costs.

The figure below shows the constraint graph of a Fuzzy CSP where $V = \{x, y, z\}$, $D = \{a, b\}$ and $C = \{c_x, c_y, c_z, c_{xy}, c_{yz}\}$. Each node models a variable and each arc models a binary constraint, while unary constraints define variables' domains. For example, c_y associates preference value 0.4 to $y = a$ and 0.7 to $y = b$.



Given an assignment s to all the variables of a soft CSP P, its preference, written $pref(P, s)$, is obtained by combining the preferences associated by each constraint to the subtuples of s referring to the variables of the constraint. For example, in fuzzy CSPs, the preference of a complete assignment is the minimum preference given by the constraints. In weighted constraints, it is the sum of the costs given by the constraints. An *optimal solution* of a soft CSP P is then a complete assignment s such that there is no other complete assignment s' with $pref(P, s) < pref(P, s')$, where $<$ is the preference ordering of the considered preference structure.

In general, finding an optimal solution for a hard or a soft CSP is computationally hard. However, it is polynomial for some classes of (soft) constraints. This is the case for tree-shaped fuzzy CSPs, where a technique called directional arc-consistency, applied bottom-up on the tree shape of the problem, is enough to make the search for an optimal solution backtrack-free and thus polynomial. A tree-shaped soft CSP is a soft CSP whose constraint graph (where nodes represent variables and arcs connect variables involved in the same constraint) is a tree.

CP-nets

While hard and soft constraints exploit preference structures to state the preference value of a value combination, other ways to model preferences are more qualitative, such as 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 $\alpha \succ \beta$) iff 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) $(a \wedge b) : c \succ \overline{c}$ would be replaced by the constraint $(a \wedge b) \Rightarrow c$.

Ethical Theories via Preference Frameworks

Many ethical theories have been defined and are used to model human behaviour when deciding what actions to take. A deontological approach to ethics involves tagging each action as either permissible, impermissible and obligatory. Given the notions defined in the previous section, it is clear that hard constraints appear to be ideal for modeling deontological ethics as defined by Kant. One could envision defining constraint problems where the actions under consideration are complete assignments to a set of decision variables modelling their different aspects and components. The constraints would be modelling ethical restrictions. Then, an action would be defined permissible if it is one of several solutions to the constraint problem, impermissible if it is not a solution, and obligatory if it is the only solution.

The criteria that Kant uses to map actions into one of the three categories are far from being defined as forbidden simultaneous assignments to some set of variables. One possible way to overcome this may be to have several hard constraints problems modelling ethical requirements in different specific domains.

Soft constraints also have many appealing properties in terms of what may be desired for modelling ethical requirements. First of all, any partial order can be represented. This is not true for other models, such as for examples, CP-nets. This is important in this context because ruling out some orderings may mean that the model may not be able to represent the "true" ethical ordering but only an approximation. Another interesting feature of soft constraints is that different combination operators can be chosen in order to aggregate preferences from different constraints. This can be useful if different ethical theories want to be modelled.

Weighted constraints appear the natural choice when it comes to model utilitarianism, that aims at maximising utilities. In fact, it easy to translate the principle of maximizing utilities to that of minimizing costs. On the other hand, fuzzy preferences which are aggregated with min well represent the fact that a violation of "ethical" constraints on any component should affect the quality of the entire option. The fundamental question is what is the set of properties of a preference aggregator which makes it suitable for handling ethical requirements? Some may be obvious, for example commutativity. Others may be a point of discussion, such as for example the fact that the aggregation of two ethical preferences cannot be "more ethical". This is called intensive property in soft constraints.

While soft constraints' quantitative approach to preferences may be appealing to model some theories, there are others which cannot be easily quantified. These are called ordinal theories in (MacAskill 2014). For such theories qualitative preferences as those modeled in CP-nets may be a better option.

Several properties of CP-nets look appealing for the objective considered in this paper. Firs of all, being able to model conditional statements may be desirable. While one may argue that ethical principles should be absolute, and not context dependent, the study of several dilemmas, such as the trolley problem, have shown that what humans regard as ethical may very well be dependent on the context and sometimes for not a very clear or rational reason.

It is reasonable to assume that artificial agents will be subject to much harsher scrutiny from an ethical stand point than humans. It is thus fundamental to be able to model what humans (or maybe a majority of humans) will consider ethical in artificial agents.

CP-nets also have the quality of not requiring numbers to express preferences. It has been argued that numbers may be a cumbersome and tedious way of representing even mundane preferences. When it comes to ethical requirements this argument may become even stronger.

One issue concerning CP-nets that will need to be addressed, is that, as mentioned above, some orderings may not be represented. Furthermore, given two options understanding if one is more desirable has a very high computational complexity. This may be unacceptable in situations where the agent is confronted by a dilemma involving two options, both with some catastrophic effect, and a decision must be made in a short amount of time.

Meta-preferences and distance over compact structures

In a social context, individual preferences are transformed little by little by incorporating elements from the societal interaction with other members of the group. This is often called "reconciliation" of individual preferences with social reason, and takes place in the context of collective choice.

To be able to describe the dynamic moving from one preference ordering to the next one (over time), and to make sure that the latter preference orderings are indeed better in terms of morality, one needs to have a way to judge preferences according to some notion of good and bad (in any of the above mentioned ethical theories). Indeed, Sen (Sen 1974) claims that morality requires judgement among preferences. To account for this, he introduced the notion of metaranking (that is, preferences over preferences) which enables to formalise individual preference modifications. A moral code could then be defined as ranking of preference rankings. That is, the moral code is defined by a structure that, by employing notions such as distance, is able to rank preferences according to their morality level. The distance intrinsic in the moral code can then be useful in measuring the deviation of any social or individual action from the moral code itself.

This approach to morality is appealing from a computational point of view. If we intend to use compact preferences models we must address two key points regarding compactly represented preferences, namely, (1) how to dynamically change them and, (2) how to define a notion of distance among them.

The first challenge has been partially addressed in the literature. Indeed, changing preferences can be seen as a form of preference elicitation or learning. This has been shown to pose some computational challenges for CP-nets (Chevaleyre et al. 2010) and has only partially been studied in the case of soft constraints (Rossi and Sperduti 1998; Khatib et al. 2007). The task of dynamically updating has also been studied in CP-nets (Cornelio et al. 2013).

Another possibility is seeing learning moral preferences as resolving uncertainty concerning what is moral. This could be represented, for example, by an extension of CP-nets called PCP-nets where preferences are expressed by a probability distribution over ordering rather than by a single ordering (Cornelio et al. 2015). Then learning can be modelled as a change in the probability distribution which lead to one in which there is no uncertainty (i.e. where one ordering has probability 1).

The second challenge is to define distances over compact preference structures. The meta-rankings defined by Sen as orderings of orderings, would be, in our case, orderings over CP-nets, where the ordering would be induced by the distance of the CP-nets from a reference "moral CP-net". To the best of our knowledge, this point has not yet been adequately explored. Defining a meaningful distance over a compact representation requires understanding the relation between that distance and the distance of the induced orderings over outcomes. Such a relation is not trivial as small differences in the compact structures can result in a large number of inversions in the induced orderings. This is true both for CP-nets and soft constraints. For CP-nets, it has been shown in (Domshlak et al. 2006) that the position of a feature in the topology of the dependency graph determines, to some extent, the magnitude of the effect of changes in its CP-table on the induced ordering. This is partially exploited in the definition of a distance over CP-nets which is proposed in (Wang et al. 2010). However, in that work the authors do not address the relation between the distance over CP-nets and their induced orderings and ignore orderings induced via transitive closure.

As far as we know, a distance on soft constraints has not been formally defined. Due to its quantitative nature, one point to clarify is if the actual values of the preference structure should matter or if only the relative ordering should count. Another subject of study should be the impact of the combination operator on the relation between the distance over compact and induced preferences.

Conclusion

As artificial agents become more intelligent they will be required to adhere to ethical requirements. In this paper we propose compact presence models as a way to embed ethical requirements and moral preferences into intelligent systems. We highlight some features of such models which make them particular appealing for this purpose and we also outline possible challenges which will have to be considered.

Acknowledgements

This work is partially supported by the project "Safety constraints and ethical principles in collective decision making systems" funded by the Future of Life Institute.

References

Allen, C.; Varner, G.; and Zinser, J. 2000. Prolegomena to any future artificial moral agent. *J. Exp. Theor. Artif. Intell.* 12(3):251–261.

Boutilier, C.; Brafman, R.; Domshlak, C.; Hoos, H.; and Poole, D. 2004. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research* 21:135–191.

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

Chevaleyre, Y.; Koriche, F.; Lang, J.; Mengin, J.; and Zanuttini, B. 2010. Learning ordinal preferences on multiattribute domains: The case of cp-nets. In *Preference Learning.* 273–296.

Cornelio, C.; Goldsmith, J.; Mattei, N.; Rossi, F.; and Venable, K. B. 2013. Updates and uncertainty in cp-nets. In *AI 2013: Advances in Artificial Intelligence - 26th Australasian Joint Conference, Dunedin, New Zealand, December 1-6, 2013. Proceedings*, 301–312.

Cornelio, C.; Grandi, U.; Goldsmith, J.; Mattei, N.; Rossi, F.; and Venable, K. B. 2015. Reasoning with pcp-nets in a multi-agent context. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, 969–977.

Dechter, R. 2003. Constraint Processing.

Domshlak, C.; Prestwich, S. D.; Rossi, F.; Venable, K. B.; and Walsh, T. 2006. Hard and soft constraints for reasoning about qualitative conditional preferences. *J. Heuristics* 12(4-5):263–285.

Khatib, L.; Morris, P.; Morris, R.; Rossi, F.; Sperduti, A.; and Venable, K. 2007. Solving and learning a tractable class of soft temporal constraints: Theoretical and experimental results. *AI Commun.* 20(3):181–209.

MacAskill, W. 2014. *Normative Uncertainty*. Ph.D. Dissertation, University of Oxford.

Rossi, F., and Sperduti, A. 1998. Learning solution preferences in constraint problems. *J. Exp. Theor. Artif. Intell.* 10(1):103–116.

Rossi, F.; van Beek, P.; and Walsh, T., eds. 2006. *Handbook of Constraint Programming*, volume 2 of *Foundations of Artificial Intelligence*. Elsevier.

Sen, A. 1974. *Choice, Ordering and Morality*. Oxford: Blackwell.

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

Wang, H.; Zhang, J.; Wan, C.; Shao, S.; Cohen, R.; Xu, J.; and Li, P. 2010. Web service selection for multiple agents with incomplete preferences. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Toronto, Canada, August 31 - September 3, 2010, Main Conference Proceedings*, 565–572.