Embedding Ethical Principles in Collective Decision Support Systems

Joshua Greene (Harvard University, USA),
Francesca Rossi (University of Padova, Italy and IBM T.J. Watson, USA),
John Tasioulas (King's College London, UK),
Kristen Brent Venable (Tulane University and IMHC, USA),
Brian Williams (MIT, USA)

Abstract

The future will see autonomous machines acting in the same environment as humans, in areas as diverse as driving, assistive technology, and health care. Think of self-driving cars, companion robots, and medical diagnosis support systems. We also believe that humans and machines will often need to work together and agree on common decisions. Thus hybrid collective decision making systems will be in great need.

In this scenario, both machines and collective decision making systems should follow some form of moral values and ethical principles (appropriate to where they will act but always aligned to humans'), as well as safety constraints. In fact, humans would accept and trust more machines that behave as ethically as other humans in the same environment. Also, these principles would make it easier for machines to determine their actions and explain their behavior in terms understandable by humans. Moreover, often machines and humans will need to make decisions together, either through consensus or by reaching a compromise. This would be facilitated by shared moral values and ethical principles.

Introduction

We believe it is important to study the embedding of safety constraints, moral values, and ethical principles in agents, within the context of collective decision making systems in societies of agents and humans.

Collective decision making involves a collection of agents who express their preferences over a shared set of possible outcomes, and a preference aggregation rule which chooses one of the options to best satisfy the agents' preferences. However, aggregating just preferences may lead to outcomes that do not follow any ethical principles or safety constraints. To embed such principles/constraints in a collective decision making system, we need to understand how to model them, how to reason with them at the level of a single agent, and how to embed them into collective decision making.

Just like individual humans, each agent that operates in a multi-agent context needs to be have an internal representation of moral values and ethical principles, as well as an ethical reasoning engine. Otherwise it would not able to explain its behaviour to others.

We claim that there is a need to adapt current logic-based modelling and reasoning frameworks, such as soft constraints, CP-nets, and constraint-based scheduling under uncertainty, to model safety constraints, moral values, and ethical principles. More precisely, we study how logic-based preference modelling frameworks can be adapted to model both (explicit) ethical principles and (implicit) moral values, as sophisticated constraints over possible actions. The constraints may be unconditional ("hard") constraints, or soft, overridable if the consequences of an individual bad action can still lead to overall good. We propose to replace preference aggregation with an appropriately developed value/ethics/preference *fusion*, an operation designed to ensure that agents' preferences are consistent with their moral values and do not override ethical principles

For ethical principles, we use hard constraints specifying the basic ethical "laws", plus some form of common-sense morality expressed as sophisticated prioritised and possibly context-dependent constraints over possible actions, equipped with a conflict resolution engine. To avoid reckless behavior in the face of uncertainty, we proposed to bound the risk of violating these ethical laws in the form of chance constraints, and we propose to develop stochastic constraint solvers that propose solutions that respect these risk bounds, based on models of environmental uncertainty. We also propose to replace preference aggregation with an appropriately developed constraint/value/ethics/preference *fusion*, an operation designed to ensure that agents' preferences are consistent with the system's safety constraints, the agents' moral values, and the ethical principles. We will leverage previous experience in developing single and multi-agent preference/constraint reasoning engines.

Today, techniques exist to enable agents to make decisions, such as scheduling activities, while satisfying some safety concerns, e.g. by using techniques from constraint-based optimization. For instance, in many critical scenarios, such as space missions where a malfunction can endanger the whole mission, activities are scheduled in such a way to maximise robustness against possible problems. We believe that these techniques can provide an inspiration to handle ethical concerns. However, we think that a much more explicit model and reasoning engine for ethical principles and moral values is needed in order to deal with them satisfactorily and allow them to evolve over time.

Which ethical principles for intelligent agents?

An intelligent agent should have capability to autonomously make good decisions, based on available data and preferences, even in the context of uncertainty, missing or noisy information, as well as incorrect input, and should be able to learn from past experience or from available historical data. Even more importantly, intelligent agents should have the ability to interact with humans, make decisions together with them, and achieve goals by working together.

An agent with these capabilities poses several crucial ethical questions. Ethical principles guide humans' behaviour. They tell us what is regarded as right or wrong. They come from values that we regards as absolute, guiding our whole life. If we want intelligent agents to enhance human capabilities, or to collaborate with humans, or even just to live and act in the same society, we need to embed in them some ethical guidelines, so they can act in their environment following values that are aligned to the human ones. Or maybe we need different values and ethical principles for agents, since they are inherently different from humans?

As Issac Asimov famously illustrated in his *I, Robot* series, explicitly programming ethical behavior is surprisingly challenging. Moral philosophy – the field that has studied explicit ethical principles most extensively – suggests three general approaches, corresponding to the three major schools of Western moral thought.

The *deontological* approach (most closely associated with Immanuel Kant) regards morality as a system of *rights* and *duties*. Here the focus is on categories of *actions*, where different actions are deemed impermissible, permissible, or obligatory based on a set of explicit rules.

The *consequentialist* approach (most closely associated with Jeremy Bentham and John Stuart Mill) aims to produce the best aggregate consequences (minimizing costs and maximizing benefits) according to a pre-specified value function. For example, a classical utilitarian approach aims to maximize the total amount of happiness.

The *virtue*- or *character*-based approach (most closely associated with Aristotle) regards ethical behavior as the product of an acquired set of behavioral dispositions that cannot be adequately summarized as an adherence to a set of deontological rules (concerning actions) or to as a commitment to maximizing good consequences.

These three approaches are well known and have been the starting point for nearly all discussions of machine ethics (Moor 1985; Bostrom 2014; Wallach and Allen 2008). Each approach has limitations that are well known. Deontological principles are easily to implement but may be rigid. Consequentialist principles require complex calculations that may be faulty. Virtue is opaque and requires extensive training with an unknown teaching criterion. There is, however, a more general problem faced by all three approaches, which is that implementing them may depend on solving daunting, general computation problems that have not been solved and may not be solved for some time.

For example, a "simple" deontological rule such as "don't lie" or "dont kill" is not specified in terms of machine movements. Rather, the machine must understand which acts of communication would constitute lying and which body movements would constitute killing in a given context. A consequentialist system would require a machine to represent all of the actions available to it, and a virtue based system would have to recognize the present situation as one with a variety of features that, together, call for one action rather than another. In other words, all three approaches, when fully implemented, seem to require something like general intelligence, which would enable the machine to represent its current situation in rich conceptual terms. Indeed, this speculation is consistent with recent research on the cognitive neuroscience of moral judgment indicating that moral judgment depends on a variety of neural systems that are not specifically dedicated to moral judgment (Greene 2014). This includes systems that enable the general representation of value and the motivation of its pursuit, visual imagery, cognitive control, and the representation of complex semantic representations. Unfortunately for Commander Data, humans have no "ethical subroutine". Real human moral judgment uses the whole brain.

What, then, can be done? Here, the human brain may nevertheless offer some guidance (Shenhav and Greene 2014). Is it morally acceptable to push someone off of a footbridge in order to save five lives (Thomson 1985)? A simple deontological response says no ("Dont kill"). A simple consequentialist response says yes ("Save the most lives"), and most humans are at least somewhat conflicted about this, but err on the side of the deontological response (in this particular case). We now know that the deontological response depends on a classically emotional neural structure known as the amygdala (reflecting emotional salience) and that the application of the consequentialist maximizing principle depends on a classically "cognitive" structure known as the dorsolateral prefrontal cortex. It seems that healthy humans engage both responses and that there is a higher-order evaluation process that depends on the ventromedial prefrontal cortex, a structure that across domains attaches emotional weight to decision variables. In other words, the brain seems to make both types of judgment (deontological and consequentialist) and then makes a higher order judgment about which lower-order judgment to trust, which may be viewed as a kind of wisdom (reflecting virtue or good character).

Such a hierarchical decision system might be implemented within an agent, or across agents. For example, some agents may apply simple rules based on action features. Others may attempt to make "limited" cost-benefit calculations. And collectively, the behavior of these agents may be determined by a weighting of these distinct, lower-level evaluative responses. Such as system might begin by following simple deontological rules, but then, either acquire more complex rules through learning, or learn when it can and cannot trust its own cost-benefit calculations. Starting with action-based rules and simple cost-benefit calculations substantially reduces the space of possible responses. Learning to trade-off between these two approaches adds some flexibility, but without requiring intractable cost-benefit calculations or lifelong moral education.

We offer this approach as just one example strategy. Of course, if we knew how we were going to solve this problem, there would be no need to bring together people with diverse

expertise. What we wish to convey is twofold: First, that we are aware of the scope of the challenge and the strengths and limitations of the extant strategies. Second, that we have some preliminary ideas for hybrid approaches that leverage insights from human moral cognition.

Another important aspect of our approach would be to consider the extent to which morality could be reduced to a set of rules that is capable of being applied in a fairly straightforward way to guide conduct , e.g. 'Do not kill', 'Keep one's promises', 'Help those in need', etc. We already know that much of common sense morality is codifiable in this way, thanks to the example of the law.

However, even if we could achieve an adequate codification of ordinary moral consciousness, at least within some domain, problems would arise. Two cases are especially worth highlighting: (a) cases where the strict application of a given rule generates an unacceptable outcome, often but not always characterisable as such by reference to some other rule that has been violated in adhering to the first, and (b) cases where the strict application of the given set of rules is unhelpfully 'silent' on the problem at hand, because it involved circumstances not foreseen by the rules.

Both phenomena (a) and (b) raise the question of when and how the strict application of a rule needs to be modified or supplemented to resolve the problem of perverse results or gaps. One important source of thinking about these issues is Aristotle's discussion of justice and equity in the Nicomachean Ethics. According to Aristotle, the common sense morality codified in law, although capable of being a generally good guide to action, will nonetheless on occasion breakdown along the lines of (a) and (b). For Aristotle, this means that the virtuous judge will need to possess, in addition to a propensity to follow legal rules, the virtue of equity. This enables the judge to use their independent judgment to correct or supplement the strict application of legal rules in cases of type (a) or (b). A key topic involves the clarification of the notion of equity, with its rule and judgment structure, as a prelude to a consideration of how this might be embedded in autonomous agents.

Designing ethical agents

No matter which approach we will choose to express ethical principles and moral values in intelligent agents, we need to find a suitable way to model it in computational terms, which is expressive enough to be able to represent all we have in mind in its full generality, and which can be reasoned upon with computational efficiency.

Ethical principles may seem very similar to the concepts of constraints (Rossi, Van Beek, and Walsh 2006; Dechter 2003) and preferences (Rossi, Venable, and Walsh 2011), which have already received a large attention in the AI literature. Indeed, constraints and preferences are a common feature of everyday decision making. They are, therefore, an essential ingredient in many reasoning tools. In an intelligent agent, we need to specify what is not allowed according to the principles, thus some form of constraints, as well as some way to prioritise among different principles, that some form of preference.

Representing and reasoning about preferences is an area of increasing theoretical and practical interest in AI. Preferences and constraints occur in real-life problems in many forms. Intuitively, constraints are restrictions on the possible scenarios: for a scenario to be feasible, all constraints must be satisfied. For example, if we have an ethical rule that says we should not kill anybody, all scenarios where people are killed are not allowed. Preferences, on the other hand, express desires, satisfaction levels, rejection degrees, or costs. For example, we may prefer an action that solves reasonably well all medical issues in a patient, rather than another one that solves completely one of them but does not address the other ones. Moreover, in many real-life optimization problems, we may have both constraints and preferences.

Preferences and constraints are closely related notions, since preferences can be seen as a form of "relaxed" constraints. For this reason, there are several constraint-based preference modeling frameworks in the AI literature. One of the most general of such frameworks defines a notion of *soft* constraints (Meseguer, Rossi, and Schiex 2006), which extends the classical constraint formalism to model preferences in a quantitative way, by expressing several degrees of satisfaction that can be either totally or partially ordered. The term soft constraints is used to distinguish this kind of constraints from the classical ones, that are usually called *hard* constraint. However, hard constraints can be seen as an instance of the concept of soft constraints where there are just two levels of satisfaction. In fact, a hard constraint can only be satisfied or violated, while a soft constraint can be satisfied at several levels.When there are both levels of satisfaction and levels of rejection, preferences are usually called bipolar, and they can be modeled by extending the soft constraint formalism (Bistarelli et al. 2006).

Preferences can also be modeled in a qualitative (also called *ordinal*) way, that is, by pairwise comparisons. In this case, soft constraints (or their extensions) are not suitable. However, other AI preference formalisms are able to express preferences qualitatively, such as CP-nets (Boutilier et al. 2004). More precisely, CP-nets provide an intuitive way to specify conditional preference statements that state the preferences over the instances of a certain feature, possibly depending on some other features. For example, we may say that we prefer driving slow to driving fast if we are in a country road. CP-nets and soft constraints can be combined, providing a single environment where both qualitative and quantitative preferences can be modeled and handled. Specific types of preferences come with their own reasoning methods. For example, temporal preferences are quantitative preferences that pertain to the position and duration of events in time. Soft constraints can be embedded naturally in a temporal constraint framework to handle this kind of preference.

An intuitive way to express preferences consists of providing a set of goals, each of which is a propositional formula, possibly adding also extra information such as priorities or weights. Candidates in this setting are variable assignments, which may satisfy or violate each goal. A weighted goal is a propositional logic formula plus a real-valued weight. The utility of a candidate is then computed

by collecting the weights of satisfied and violated goals, and then aggregating them. Often only violated goals count, and their utilities are aggregated with functions such as sum or maximin. In other cases, we may sum the weights of the satisfied goals, or we may take their maximum weight. Any restriction we may impose on the goals or the weights, and any choice of an aggregation function, give a different language. Such languages may have drastically different properties in terms of their expressivity, succinctness, and computational complexity.

In the quantitative direction typical of soft constraints, there are also other frameworks to model preferences. The most widely used assumes we have some form of independence among variables, such as mutual preferential independence. Preferences can then be represented by an additive utility function in deterministic decision making, or utility independence, which assures an additive representation for general scenarios. However, this assumption often does not hold in practice since there is usually some interaction among the variables. To account for this, models based on interdependent value additivity have been defined which allows for some interaction between the variables while preserving some decomposability. This notion of independence, also called generalized additive independence (GAI), allows for the definition of utility functions which take the form of a sum of utilities over subsets of the variables. GAI decompositions can be represented by a graphical structure, called a GAI net, which models the interaction among variables, and it is similar to the dependency graph of a CP-net or to the junction graph of a Bayesian network. GAI decompositions have been used to provide CP-nets with utility functions, obtaining the so-called UCP networks.

Preferences and ethical principles in collective decision making systems

If agents and humans will be part of a hybrid collective decision making system, and thus will make collective decisions, based on their preferences over the possible outcomes, can ethical principles for such decision system be modelled just like the preferences of another dummy agent, or should they be represented and treated differently? Are the knowledge representation formalisms that are usually used in AI to model preferences suitable to model values as well, or should we use something completely different? A very simple form of values could be modelled by constraints, so that only feasible outcomes can be the results of a collective decisions process. But values and ethical principles could often take a graded form, thus resembling a kind of preference. Also, should individual and collective ethical principles be modelled differently?

We believe that some of the answers to these questions may exploit the existing literature on preference aggregation (Rossi, Venable, and Walsh 2011). Indeed, an important aspect of reasoning about preferences is preference aggregation. In multi-agent systems, we often need to combine the preferences of several agents. More precisely, preferences are often used in collective decision making when multiple agents need to choose one out of a set of possible decisions: each agent expresses its preferences over the possible decisions, and a centralized system aggregates such preferences to determine the "winning" decision. Preferences are also the subject of study in social choice, especially in the area of elections and voting theory (Arrow and amd K. Suzumara 2002). In an election, the voters express their preferences over the candidates and a voting rule is used to elect the winning candidate. Economists, political theorist, mathematicians, as well as philosophers have invested considerable effort in studying this scenario and have obtained many theoretical results about the desirable properties of the voting rules that one can use.

Since the voting setting is closely related to multi-agent decision making, in recent years the area of multi-agent systems has witnessed a growing interest in trying to reuse social choice results in the multi-agent setting. However, it soon became clear that an adaptation of such results is necessary, since several issues, which are typical of multi-agent settings and AI scenarios, usually do not occur, or have a smaller impact, in typical voting situations. In a multi-agent system, the set of candidates can be very large with respect to the set of voters. Usually in social choice it is the opposite: there are many voters and a small number of candidates. Also, in many AI scenarios, the candidates often have a combinatorial structure. That is, they are defined via a combination of features. Moreover, the preferences over the features are often dependent on each other. In social choice, usually the candidates are tokens with no structure. In addition, for multi-issue elections, the issues are usually independent of each other. This combinatorial structure allows for the compact modelling of the preferences over the candidates. Therefore, several formalisms have been developed in AI to model such preference orderings. In social choice, little emphasis is put on how to model preferences, since there are few candidates, so one can usually explicitly specify a linear order. In AI, a preference ordering is not necessarily linear, but it may include indifference and incomparability. Moreover, often uncertainty is present, for example in the form of missing or imprecise preferences. In social choice, usually all preferences are assumed to be present, and a preference order over all the candidates is a linear order that is explicitly given as a list of candidates. Finally, multi-agent systems must consider the computational properties of the system. In social choice this usually has not been not a crucial issue.

It is therefore very interesting to study how social choice and AI can fruitfully cooperate to give innovative and improved solutions to aggregating preferences of multiple agents. In our effort, since we intend to deal with ethical issues in collective decision making, we need to understand what modifications to the usual preference aggregation scenario should be done to account for them, and how they can be handled satisfactorily when making collective decisions. Collective decision making in the presence of feasibility constraints is starting to be considered in the literature (Grandi et al. 2014). However, ethical principles and safety constraints will be much more complex than just a set of constraints, so we need to understand the computational and expressiveness issues arising in this scenario.

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

Arrow, K. J., and amd K. Suzumara, A. K. S. 2002. *Handbook of Social Choice and Welfare.* North-Holland, Elsevier.

Bistarelli, S.; Pini, M. S.; Rossi, F.; and Venable, K. B. 2006. Bipolar preference problems: Framework, properties and solving techniques. In *Recent Advances in Constraints (CSCLP 2006)*, volume 4651 of *LNCS*, 78–92. Springer.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Boutilier, C.; Brafman, R. I.; Domshlak, C.; Hoos, H. H.; and Poole, D. 2004. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *J. Artif. Intell. Res. (JAIR)* 21:135–191.

Dechter, R. 2003. *Constraint Processing*. Morgan Kaufmann.

Grandi, U.; Luo, H.; Maudet, N.; and Rossi, F. 2014. Aggregating cp-nets with unfeasible outcomes. In *Principles and Practice of Constraint Programming - 20th International Conference, CP 2014, Lyon, France, September 8-12, 2014. Proceedings*, 366–381.

Greene, J. D. 2014. *The cognitive neuroscience of moral judgment and decision-making*. MIT Press.

Meseguer, P.; Rossi, F.; and Schiex, T. 2006. Soft constraints. In *Handbook of constraint programming*. Elsevier. chapter 9, 281–328.

Moor, J. H. 1985. What is computer ethics? *Metaphilosophy* 16(4):266–275.

Rossi, F.; Van Beek, P.; and Walsh, T., eds. 2006. *Handbook of Constraint Programming*. Elsevier.

Rossi, F.; Venable, K. B.; and Walsh, T. 2011. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Shenhav, A., and Greene, J. D. 2014. Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience* 34(13):4741–4749.

Thomson, J. J. 1985. The trolley problem. *Yale Law Journal* 94:1395.

Wallach, W., and Allen, C. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.