Moral preferences

Francesca Rossi
IBM T.J. Watson Research center*
frossi@it.ibm.com

1 Motivation and Introduction

How do humans or machines make a decision? Whenever we make a decision, we consider our preferences over the possible options. Also, in a social context, collective decisions are made by aggregating the preferences of the individuals. AI systems that support individual and collective decision making have been studied for a long time, and several preference modelling and reasoning frameworks have been defined and exploited in order to provide rationality to the decision process and its result.

However, little effort has been devoted to understand whether this decision process, or its result, is ethical or moral. Rationality does not imply morality. How can we embed morality into a decision process? And how do we ensure that the decision we make, as an individual or a collectivity of individuals, are moral? In other words, how do we pass from the individuals' personal preferences to moral behaviour and decision making?

When we pass from humans to AI systems, the task of modelling and embedding morality and ethical principles is even more vague and elusive. Are the existing ethical theories applicable also to AI systems? On one hand, things seem easier since we can narrow the scope of an AI system, so that the contextual information can help us in define the correct moral values it should work according to. However, it is not clear what moral values we should embed in the system, nor how to embed them. Should we code them in a set of rules, or should we let the system learn the values by observing us humans?

Preferences and ethical theories are not that different in one respect: they both define priorities over actions. So, can we use existing preference formalisms to also model ethical theories? We discuss how to exploit and adapt current preference formalisms in order to model morality and ethics theories, as well as the dynamic integration of moral code into personal preferences. We also discuss the use of meta-preferences, since morality seems to need a way to judge preferences according to their morality level.

It is imperative that we build intelligent systems which behave morally. To work and live with us, we need to trust such systems, and this requires that we are "reasonably" sure that it behaves morally, according to values that are aligned to the human ones. Otherwise, we would not let a robot take care of our elderly people or our kids, nor a car to drive for us, nor we would listen to a decision support system in any healthcare scenario. Of course the word "reasonable" makes sense when the application domain does not include critical situations (like suggesting a friend on a social media or a movie in an online selling system). But when the AI system is helping (or replacing) humans in critical domains such as healthcare, then we need to have a guarantee that nothing morally wrong will be done.

In this extended abstract we introduce some issues in embedding morality into intelligent systems. A few research questions are defined, with no answer to them, with the hope that the discussion raised by the questions will shed some light onto the possible answers.

2 Preference modelling and reasoning

Preferences have been studied for a long time in AI, both in the area of knowledge representation and in multi-agent systems. Several frameworks have been defined to model different kinds of preferences, such as qualitative (as in, e.g., "I prefer blue to red") and quantitative ones (as in, e.g., "I give 5 stars to Breakfast at Tiffany's and 2 stars to Terminator"). In general preferences are defining an ordering over a set of options. This order can be total and strict, but in practice it may have a lot of ties and incomparability.

When the set of options is very large, and each option is defined by a set of features (such as a car, which can be defined by it model, its colour, its engine, etc.), preferences can be expressed over single features of small sets of them, rather than entire options (as in, e.g., "If I buy a convertible, I prefer it to be red rather than white"). This allows for a faster and easier preference specification phase, as well as for more efficient preference elicitation. Several ways have been defined to pass from such compact ways to model preferences over features to the preference ordering over the options. However, it is possible to reason about such preferences without generating the exponentially large ordering over the options, which makes preferences reasoning tractable in some cases. Examples of framework to do this are constraints [Rossi *et al.*, 2006], soft constraints [Meseguer *et al.*, 2005] and CP-nets [Boutilier *et al.*, 2004].

Once an individual's preferences over the possible options are specified, we need to be able to find the most preferred

*On leave from University of Padova, Italy

option, or the next best option, or to compare two options that may be presented to us. Several algorithms to performs such tasks have been defined [Brafman *et al.*, 2010; Boutilier *et al.*, 2004].

When individuals, or AI systems, are part of a social environment and need to make collective decisions, individual's preferences are aggregated (for example via some voting rule) and an option is chosen for the whole group. Many voting rules have been defined and studied, as well as their properties [Arrow *et al.*, 2002]. Issues such as manipulation, control, bribery, as well as properties such as fairness and unanimity have long being investigated, in order to define decision support systems that behave as desired [Airiau *et al.*, 2011; Fargier *et al.*, 2012; Conitzer *et al.*, 2011; Xia and Conitzer, 2010; Lang *et al.*, 2007; Pini *et al.*, 2011; Pozza *et al.*, 2011; Gonzales *et al.*, 2008; Maran *et al.*, 2013; Purrington and Durfee, 2007; Lang and Xia, 2009].

3 From preferences to morality

To trust an AI system, like a companion robot or a self-driving car, we need to be reasonably sure that it behaves morally, according to values that are aligned to the human ones. Otherwise, we would not let a robot take care of our elderly people or our kids, nor a car to drive for us, nor we would listen to a decision support system in any healthcare scenario. So it is imperative that we understand how to provide AI systems with morality [Musschenga and van Harskamp, 2013; Wallach and Allen, 2009; Greene *et al.*, 2016].

Morality and ethical behaviour are based on prioritising actions on the basis of what is morally right or wrong. Many ethical theories have been defined and studied in the psychology literature. They include the following ones:

- Consequentialism: Action consequences are evaluated interns of a scale of good and bad, and an agent should choose the action that minimise the bad and maximises the good.
- Virtue Ethics: An agent should choose actions that satisfy some pre-defined set of virtues
- Deontologism: Actions are predefined as good or bad, and an agent should choose the best action, no matter the consequences.

No matter which ethical theory one decides to use, the notion of right and wrong of course depends on the context in which humans (or machines) function, so formally an ethical theory can be defined as a function from a context to a partial ordering over actions. Indeed, usually we have a partial order over actions, since some actions could be incomparable to others. As one may notice by looking at the previous section on preferences, this is not that different from what preferences define: a partial order over possible options (of actions, or decisions in general). So it makes sense to investigate the possible use of preference frameworks in modelling and embedding morality into AI systems.

Research question 1: Are existing preference modelling and reasoning frameworks ready to be used also to model and reason with ethical principles and moral code, or we need to adapt them or invent new ones?

If we had the "moral" partial order and the "preference" partial order for each individual, one could try to merge them in some way, to obtain a "moral preference ordering". For example, two CP-nets modelling the moral and the preference orderings could be syntactically or semantically merged via operators that could give priority to the moral CP-net and let the preference one dictate the behaviour only when it is not in conflict with the moral one. The technical details have not been spelled out yet, but one could imagine several reasonable ways of doing this.

Research question 2: Given a moral and an ethical ordering over actions, how to combine them? Given such orderings in the forms of CP-nets or soft constraints, or other compact formalisms to model preferences, how to combine them? What properties should we desire about their combination?

However, knowing the preferences of an individual is already a difficult task. Elicitation and learning framework have bee proposed in order to do that in a way that is most faithful to the "real" preferences of the individual. Knowing the moral ordering of an individual is even more difficult. And this is even more so when we are in a social context, since this may make individuals change their moral attitudes over time because of social interaction. The existing approaches to define ethical principles in AI systems range from trying to code ethical principles in the form of rules, to letting the system "learn" such principles from a (possibly supervised) observation of the behaviour of humans in similar settings. Some AI systems try to list the set of rules to use in self-driving cars to solve ethical dilemmas like the trolley problem. However, such approaches are usually not general, since it is unfeasible to foresee all possible situations in a very wide scenario. On the other hand, other approaches use, for example, inverse reinforcement learning [Ng and Russell, 2000] to try to learn morality from human behaviour. I personally feel that the best results could be obtained by combining these two approaches, although it is not clear yet how to do it best.

Research question 3: How to combine bottom-up learning approaches with top-down rule-based approaches in defining ethical principles for AI systems?

Research question 4: Recently, the most successful AI systems are based on statistical machine learning approaches that, by their nature, do not provide a natural way to explain or justify their decisions (or suggestions), nor they assure optimality. If we employ this approach also for embedding morality into a machine, how are we going to prove that nothing morally wrong will happen?

4 Morality by meta-preferences

As mentioned above, in a social context, individual preferences are transformed little by little by incorporating reasonable 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 over the next one (in time), and to make sure that the later 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.

Research question 5: Given a moral code, in a social choice context, where individuals submit their preference ordering and the result is a collective preference ordering, how to measure the deviation of the collective ordering from a moral code? And how to measure the deviation of individuals from a collective moral code?

If an individual modifies its preference ordering from a morally low to a morally higher ordering, we should want to use collective decision making system in which such a move leads to collective actions of higher morality. That is, some form of monotonicity should be desired.

Research question 6: Which properties should be desired in a moral preference aggregation environment?

5 Morality in narrow AI systems

In [Greene, 2014] it is shown that human moral judgment doesn't come from a dedicated moral system, but it is rather the product of the interaction of many general-purpose brain networks, each working and being useful in narrow contexts. So it seems that humans need a general purpose brain in order to be moral. Is it true also for AI systems?

Research question 7: Can narrow AI systems be moral? If humans bring all of their general intelligence to bear when making moral decisions, even fairly simple ones, does that that mean that we have to solve Artificial General Intelligence in order to produce something useful?

6 Conclusions

Intelligent systems are going to be more and more pervasive in our everyday lives. To name just a few applications, they will take care of elderly people and kids, they will drive for us, and they will suggest doctors how to cure a disease. However, we cannot let them do all this very useful and beneficial tasks if we don't trust them. To build trust, we need to be sure that they act in a morally acceptable way. So it is important to understand how to embed moral values into intelligent machines.

Existing preference modelling and reasoning framework can be a starting point, since they define priorities over actions, just like an ethical theory does. However, many more issues are involved when we mix preferences (that are at the core of decision making) and morality, both at the individual level and in a social context. We have listed some of these questions, hoping that this short paper can generate some answers.

References

[Airiau *et al.*, 2011] S. Airiau, U. Endriss, U. Grandi, D. Porello, and J. Uckelman. Aggregating dependency graphs into voting agendas in multi-issue elections. In *Proceedings of IJCAI 2011*, pages 18–23, 2011.

[Arrow *et al.*, 2002] K. J. Arrow, A. K. Sen, and K. Suzumura. *Handbook of Social Choice and Welfare.* North-Holland, 2002.

[Boutilier *et al.*, 2004] C. Boutilier, R. I. Brafman, C. Domshlak, H. H. Hoos, and D. Poole. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *JAIR*, 21:135–191, 2004.

[Brafman *et al.*, 2010] R. I. Brafman, F. Rossi, D. Salvagnin, K. B. Venable, and T. Walsh. Finding the next solution in constraint- and preference-based knowledge representation formalisms. In *Proceedings of KR 2010*, 2010.

[Conitzer *et al.*, 2011] V. Conitzer, J. Lang, and L. Xia. Hypercubewise preference aggregation in multi-issue domains. In *Proceedings of IJCAI 2011*, pages 158–163, 2011.

[Fargier *et al.*, 2012] H. Fargier, J. Lang, J. Mengin, and N. Schmidt. Issue-by-issue voting: an experimental evaluation. In *Proceedings of MPREF 2012*, 2012.

[Gonzales *et al.*, 2008] C. Gonzales, P. Perny, and S. Queiroz. Preference aggregation with graphical utility models. In *Proceedings of AAAI 2008*, pages 1037–1042, 2008.

[Greene *et al.*, 2016] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. Embedding ethical principles in collective decision support systems. In *Proceedings AAAI 2016*. AAAI Press, 2016.

[Greene, 2014] Joshua Greene. The cognitive neuroscience of moral judgment and decision making. In *The Cognitive Neurosciences V (ed. M.S. Cazzaniga)*. MIT Press, 2014.

[Lang and Xia, 2009] J. Lang and L. Xia. Sequential composition of voting rules in multi-issue domains. *Mathematical social sciences*, 57:304–324, 2009.

[Lang *et al.*, 2007] J. Lang, M. S. Pini, F. Rossi, K. B. Venable, and T. Walsh. Winner determination in sequential majority voting. In *Proceedings of IJCAI 2007*, pages 1372–1377, 2007.

[Maran *et al.*, 2013] A. Maran, N. Maudet, M. S. Pini, F. Rossi, and K. B. Venable. A framework for aggregating influenced CP-nets and its resistance to bribery. In *Proceedings of AAAI 2013*, 2013.

[Meseguer *et al.*, 2005] P. Meseguer, F. Rossi, and T. Schiex. Soft constraints. In P. Van Beek F. Rossi and T. Walsh, editors, *Handbook of Constraint Programming*. Elsevier, 2005.

[Musschenga and van Harskamp, 2013] Bert Musschenga and Anton (eds.) van Harskamp. *What Makes Us Moral? On the capacities and conditions for being moral*. Springer, 2013.

[Ng and Russell, 2000] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 2000.

[Pini *et al.*, 2011] M. S. Pini, F. Rossi, K. B. Venable, and T. Walsh. Incompleteness and incomparability in preference aggregation: Complexity results. *Artif. Intell.*, 175(7-8):1272–1289, 2011.

[Pozza *et al.*, 2011] G. Dalla Pozza, M. S. Pini, F. Rossi, and K. B. Venable. Multi-agent soft constraint aggregation via sequential voting. In *Proceedings of IJCAI 2011*, pages 172–177, 2011.

[Purrington and Durfee, 2007] K. Purrington and E. H. Durfee. Making social choices from individuals' CP-nets. In *Proceedings of AAMAS 2007*, pages 1122–1124, 2007.

[Rossi *et al.*, 2006] F. Rossi, P. Van Beek, and T. Walsh. *Handbook of Constraint Programming*. Elsevier, 2006.

[Sen, 1974] Amartya Sen. Choice, ordering and morality. In *Practical Reason, Krner S. (ed)*. Oxford, 1974.

[Wallach and Allen, 2009] Wendell Wallach and Colin Allen. *Moral Machines*. Oxford, 2009.

[Xia and Conitzer, 2010] L. Xia and V. Conitzer. Strategy-proof voting rules over multi-issue domains with restricted preferences. In *Proceedings of WINE 2010*, pages 402–414, 2010.