Moral Decision Making Frameworks for Artificial Intelligence

Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, Max Kramer
Duke University, Durham, NC 27708, USA
{vincent.conitzer,walter.sinnott-armstrong,
jana.schaich.borg,yuan.deng,max.kramer}@duke.edu

Abstract

The generality of decision and game theory has enabled domain-independent progress in AI research. For example, a better algorithm for finding good policies in (PO)MDPs can be instantly used in a variety of applications. But such a general theory is lacking when it comes to *moral* decision making. For AI applications with a moral component, are we then forced to build systems based on many ad-hoc rules? In this paper we discuss possible ways to avoid this conclusion.

Introduction

As deployed AI systems become more autonomous, they increasingly face moral dilemmas. An often-used example is that of a self-driving car that faces an unavoidable accident, but has several options how to act, with different effects on its passengers and others in the scenario. (See, for example, Bonnefon *et al.* (2016).) But there are other examples where AI is already used to make decisions with life-or-death consequences. Consider, for example, kidney exchanges. These cater to patients in need of a kidney that have a willing live donor whose kidney the patient's body would reject. In this situation, the patient may be able to swap donors with another patient in the same situation. (More complex arrangements are possible as well.) For these exchanges, algorithms developed in the AI community are already used to determine which patients receive which kidneys (see, e.g., Dickerson and Sandholm (2015)). While it may be possible to find special-purpose solutions for moral decision making in these domains, in the long run there is a need for a general framework that an AI agent can use to make moral decisions in a wider variety of contexts. In this paper, we lay out some possible roadmaps for arriving at such a framework.

Motivation

Most AI research is conducted within straightforward utilitarian or consequentialist frameworks, but these simple approaches can lead to counterintuitive judgments from an ethical perspective. For example, most people consider it immoral to harvest a healthy patient's organs to save the lives of two or even five other patients. Research in ethics and moral psychology elucidates our moral intuitions in such examples by distinguishing between *doing* and *allowing*, emphasizing the role of *intent*, applying general rules about kinds of actions (such as "Don't kill"), and referring to rights (such as the patient's) and roles (such as the doctor's). Incorporating these morally relevant factors among others could enable AI to make moral decisions that are safer, more robust, more beneficial, and acceptable to a wider range of people.[1]

To be useful in the development of AI, our moral theories must provide more than vague, general criteria. They must also provide an operationalizable, and presumably quantitative, theory that specifies which particular actions are morally right or wrong in a wide range of situations. This, of course, also requires the agent to have a language in which to *represent* the structure of the actions being judged (Mikhail, 2007) and the morally relevant features of actions (Gert, 2004) along with rules about how these features interact and affect moral judgments. Moral theory and AI need to work together in this endeavor.

Multiple approaches can be taken to arrive at general-purpose procedures for automatically making moral decisions. One approach is to use game theory. Game-theoretic formalisms are widely used by artificial intelligence researchers to represent multiagent decision scenarios, but, as we will argue below, its solution concepts and possibly even its basic representation schemes need to be extended in order to provide guidance on moral behavior. Another approach is to use machine learning. We can use the moral philosophy and psychology literatures to identify features of moral dilemmas that are relevant to the moral status of possible actions described in the dilemmas. Human subjects can be asked to make moral judgments about a set of moral dilemmas in order to obtain a labeled data set. Then, we can train classifiers based on this data set and the identified features. (Compare also the top-down vs. bottom-up distinction in automated moral decision making, as described by Wallach and Allen (2008).) We will discuss these two approaches in turn.

[1] The point that, as advanced AI acquires more autonomy, it is essential to bring moral reasoning into it has been made previously by others—e.g., Moor (2006).

Examples

In this paper, we will take a very broad view of what constitutes a moral dilemma (contrast Sinnott-Armstrong (1988)). As a simple example, consider the *trust game* (Berg *et al.*, 1995). In the trust game, player 1 is given some amount of money—say, \$100. She[2] is then allowed to give any fraction of this money back to the experimenter, who will then triple this returned money and give it to player 2. Finally, player 2 may return any fraction of the money he has received to player 1. For example, player 1 might give \$50 back, so that player 2 receives $3 \cdot \$50 = \$150$, who then might give \$75 back, leaving player 1 with $\$50 + \$75 = \$125$. The most straightforward game-theoretic analysis of this game assumes that each player, at any point in the game, is interested only in maximizing the amount of money she herself receives. Under this assumption, player 2 would never have any reason to return any money to player 1. Anticipating this, player 1 would not give any money, either. However, despite this analysis, human subjects playing the trust game generally *do* give money in both roles (Berg *et al.*, 1995). One of the reasons why is likely that many people feel it is *wrong* for player 2 not to give any money back after player 1 has decided to give him some (and, when in the role of player 1, they expect player 2 not to take such a wrong action).

This case study illustrates a general feature of moral reasoning. Most people consider not only the consequences of their actions but also the *setting* in which they perform their actions. They ask whether an act would be unfair or selfish (because they are not sharing a good with someone who is equally deserving), ungrateful (because it harms someone who benefited them in the past), disloyal (by betraying a friend who has been loyal), untrustworthy (because it breaks a promise), or deserved (because the person won a competition or committed a crime). In these ways, moral reasoners typically look not only to the future but also to the past.

Of course, not everyone will agree about which factors are morally relevant, and even fewer people will agree about which factor is the most important in a given conflict. For example, some people will think that it is morally wrong to lie to protect a family member, whereas others will think that lying in such circumstances is not only permitted but required. Nonetheless, a successful moral AI system does not necessarily have to dictate one true answer in such cases. It may suffice to know how much various groups value different factors or value them differently. Then when we code moral values into AI, we would have the option of either using the moral values of a specific individual or group—a type of moral relativism—or giving the AI some type of *social-choice-theoretic* aggregate of the moral values that we have inferred (for example, by letting our models of multiple people's moral values *vote* over the relevant alternatives, or using only the moral values that are common to all of them). This approach suggests new research problems in the field of *computational social choice* (see, e.g., Brandt *et al.* (2013, 2015)). Rossi (2016) has described related, but distinct social choice problems where (not necessarily moral) preferences are either aggregated together with a moral ranking of all the alternatives, or the preferences are themselves ranked according to a moral ordering (see also Greene *et al.* (2016)).

[2]We use "she" for player 1 or a generic player, and "he" for player 2.

Abstractly Representing Moral Dilemmas: A Game-Theoretic Approach

For us humans, the most natural way to describe a moral dilemma is to use natural language. However, given the current state of AI in general and of natural language processing in particular, such verbal descriptions will not suffice for our purposes. Moral dilemmas will need to be more abstractly represented, and as is generally the case in AI research, the choice of representation scheme is extremely important. In this section, we consider an approach to this problem inspired by game theory.

Game-Theoretic Representation Schemes

Game theory (see, e.g., Fudenberg and Tirole (1991)) concerns the modeling of scenarios where multiple parties (henceforth, *agents*) have different interests but interact in the same domain. It provides various natural representation schemes for such multiagent decision problems. Scenarios described in game theory involve sequences of actions that lead to different agents being better or worse off to different degrees. Since moral concepts—such as selfishness, loyalty, trustworthiness, and fairness—often influence which action people choose to take, or at least believe they *should* take, in such situations, game theory is potentially a good fit for abstractly representing moral dilemmas.

One of the standard representation schemes in game theory is that of the *extensive form*, which is a generalization of the game trees studied in introductory AI courses. The extensive-form representation of the trust game (or rather, a version of it in which player 1 can only give multiples of \$50 and player 2 only multiples of \$100) is shown in Figure 1.

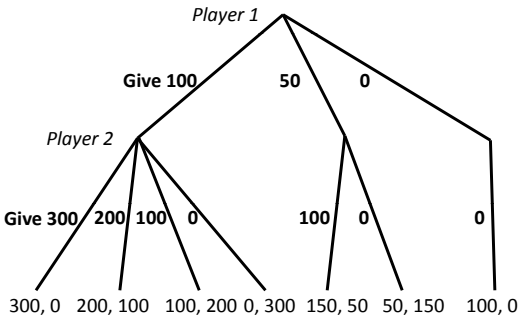

Figure 1: The trust game. Each edge corresponds to an action in the game and is labeled with that action. Each bottom (leaf) node corresponds to an outcome of the game and is labeled with the corresponding payoffs for player 1 and player 2, respectively.

We will turn to the question of whether such representation schemes suffice to model moral dilemmas more generally shortly. First, we discuss how to *solve* such games.

Moral Solution Concepts

The standard solution concepts in game theory assume that each agent pursues nothing but its own prespecified utility. If we suppose in the trust game that each player just seeks to maximize her own monetary payoff, then game theory would prescribe that the second player give nothing back regardless of how much he receives, and consequently that the first player give nothing.[3] However, this is not the behavior observed in experiments with human subjects. Games that elicit human behavior that does not match game-theoretic analyses, such as the trust game, are often used to criticize the game-theoretic model of behavior and have led to the field of *behavioral game theory* (Camerer, 2003). While in behavioral game theory, attention is often drawn to the fact that humans are not infinitely rational and cannot be expected to perform complete game-theoretic analyses in their heads, it seems that this is not the primary reason that agents behave differently in the trust game, which after all is quite simple. Rather, it seems that the simplistic game-theoretic solution fails to account for ethical considerations.

In traditional game theory's defense, it should be noted that an agent's utility may take into account the welfare of others, so it is possible for *altruism* to be captured by a game-theoretic account. However, what is morally right or wrong also seems to depend on past actions by other players. Consider, for example, the notion of *betrayal*: if another agent knowingly enables me either to act to benefit us both, or to act to benefit myself even more while significantly hurting the other agent, doing the latter seems morally wrong. This, in our view, is one of the primary things going on in the trust game. The key insight is that to model this phenomenon, we cannot simply first assess the agents' other-regarding preferences, include these in their utilities at the leaves of the game, and solve the game (as in the case of pure altruism). Rather, the analysis of the game (solving it) must be *intertwined* with the assessment of whether an agent morally should pursue another agent's well-being. This calls for novel *moral solution concepts* in game theory.

We have already done some conceptual and algorithmic work on a solution concept that takes such issues into account (Letchford *et al.*, 2008). This solution concept involves repeatedly solving the game and then modifying the agents' preferences based on the solution. The modification makes it so that (for example) player 2 wants to ensure that player 1 receives at least what she could have received in the previous solution, unless this conflicts with player 2 receiving at least as much as he would have received in the previous solution. For example, in the trust game player 2's preferences are modified so that he values player 1 receiving back at least what she gave to player 2.

What Is Left Out & Possible Extensions

The solution concept from Letchford *et al.* (2008) is defined only in very restricted settings, namely 2-player perfect-information[4] games. One research direction is to generalize the concept to games with more players and/or imperfect information. Another is to define different solution concepts that capture other ethical concerns.

Zooming out, this general approach is inherently limited by the aspects of moral dilemmas that can be captured in game-theoretic representations. While we believe that the standard representation schemes of game theory can capture much of what is relevant, they may not capture everything that is relevant. For example, in moral philosophy, a distinction is often made between *doing* harm and *allowing* harm. Consider a situation where a runaway train will surely hit and kill exactly one innocent person (player 2) standing on a track, unless player 1 intervenes and puts the train on another track instead, where it will surely hit and kill exactly one other innocent person (player 3). The natural extensive form of the game (Figure 2) is entirely symmetric and thereby cannot be used to distinguish between the two alternatives. (Note that the labels on the edges are formally not part of the game.) However, many philosophers (as well as non-philosophers) would argue that there is a significant distinction between the two alternatives, and that switching the train to the second track is morally wrong. We propose that the action-inaction distinction could be addressed by slightly extending the extensive-form representation so that at every information set (decision point), one action is labeled as the "passive" action (e.g., leaving the train alone). Other extensions may be needed as well. For example, we may take into account what each agent in the game *deserves* (according to some theory of desert), which may require us to further extend the representation scheme.[5]

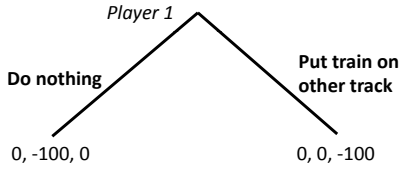

Figure 2: "Runaway train." Player 1 must choose whether to allow player 2 to be hurt or to hurt player 3 instead.

A broader issue is that in behavioral game and decision theory it is well understood that the way the problem is *framed*—i.e., the particular language in which the problem is described, or even the order in which dilemmas are presented—can significantly affect human subjects' decisions. That is, two ways of describing the same dilemma can

[3] The technical name for this type of analysis is *backward induction*, resulting in behavior that constitutes a *subgame perfect Nash equilibrium* of the game.

[4] In a *perfect-information* game, the current state is fully observable to each player (e.g., chess), in contrast to imperfect-information games (e.g., poker).

[5] Note that, to the extent the reasons for what an agent deserves are based *solely on the agent's earlier actions in the game under consideration*, solution concepts such as those described above might in fact capture this. If so, then the only cases in which we need to extend the representation scheme are those where what an agent deserves is external to the game under study (e.g., the agent is a previously convicted criminal).

produce consistently different responses from human subjects (Kahneman and Tversky, 2000). The same is surely the case for moral dilemmas (Sinnott-Armstrong, 2008). Moral AI would need to replicate this behavior if the goal is to mirror or predict human moral judgments. In contrast, if our goal is to make *coherent* moral judgments, then moral AI might instead need to avoid such framing effects.

Setting up a Machine Learning Framework

Another approach for developing procedures that automatically make moral decisions is based on *machine learning* (see, e.g., Mitchell (1997)). We can assemble a training set of moral decision problem instances labeled with human judgments of the morally correct decision(s), and allow our AI system to generalize. (Other work has focused on obtaining human judgments not of the actions themselves, but of *persuasion strategies* in such scenarios (Stock *et al.*, 2016).) To evaluate this approach with current technology, it is insufficient to represent the instances in natural language; instead, we must represent them more abstractly. What is the right representation scheme for this purpose, and what features are important? How do we construct and accurately label a good training set?

Representing Dilemmas by Their Key Moral Features

When we try to classify a given action in a given moral dilemma as morally right or wrong (as judged by a given human being), we can try to do so based on various *features* (or *attributes*) of the action. In a restricted domain, it may be relatively clear what the relevant features are. When a self-driving car must decide whether to take one action or another in an impending-crash scenario, natural features include the expected number of lives lost for each course of action, which of the people involved were at fault, etc. When allocating a kidney, natural features include the probability that the kidney is rejected by a particular patient, whether that patient needs the kidney urgently, etc. Even in these scenarios, identifying *all* the relevant features may not be easy. (E.g., is it relevant that one potential kidney recipient has made a large donation to medical research and the other has not?) However, the primary goal of a *general* framework for moral decision making is to identify abstract features that apply across domains, rather than to identify every nuanced feature that is potentially relevant to isolated scenarios. The literature in moral psychology and cognitive science may guide us in identifying these general concepts. For example, Haidt and Joseph (2004) have proposed five moral foundations—harm/care, fairness/reciprocity, loyalty, authority, and purity. Recent research has added new foundations and subdivided some of these foundations (Clifford *et al.*, 2015). The philosophy literature can similarly be helpful; e.g., Gert (2004) provides a very inclusive list of morally relevant features.

Classifying Actions as Morally Right or Wrong

Given a labeled dataset of moral dilemmas represented as lists of feature values, we can apply standard machine learning techniques to learn to classify actions as morally right or wrong. In ethics it is often seen as important not only to act in accordance with moral principles but also to be able to *explain why* one's actions are morally right (Anderson and Anderson, 2007; Bostrom and Yudkowsky, 2014); hence, *interpretability* of the resulting classifier will be important.

Of course, besides making a binary classification of an action as morally right or wrong, we may also make a quantitative assessment of *how* morally wrong the action is (for example using a regression), an assessment of *how probable* it is that the action is morally wrong (for example using a Bayesian framework), or some combination of the two. Many further complicating factors can be added to this simple initial framework.

Discussion

A machine learning approach to automating moral judgements is perhaps more flexible than a game-theoretic approach, but the two can complement each other. For example, we can apply moral game-theoretic concepts to moral dilemmas and use the output (say, "right" or "wrong" according to this concept) as one of the features in our machine learning approach. On the other hand, the outcomes of the machine learning approach can help us see which key moral aspects are missing from our moral game-theoretic concepts, which will in turn allow us to refine them.

It has been suggested that machine learning approaches to moral decisions will be limited because they will *at best* result in human-level moral decision making; they will never exceed the morality of humans. (Such a worry is raised, for example, by Chaudhuri and Vardi (2014).) But this is not necessarily so. First, aggregating the moral views of multiple humans (through a combination of machine learning and social-choice theoretic techniques) may result in a morally better system than that of any individual human, for example because idiosyncratic moral mistakes made by individual humans are washed out in the aggregate. Indeed, the learning algorithm may well decide to output a classifier that disagrees with the labels of some of the instances in the training set (see Guarini (2006) for a discussion of the importance of being able to revise initial classifications). Second, machine learning approaches may identify general principles of moral decision making that humans were not aware of before. These principles can then be used to improve our moral intuitions in general. For now, moral AI systems are in their infancy, so creating even human-level automated moral decision making would be a great accomplishment.

Conclusion

In some applications, AI systems will need to be equipped with moral reasoning capability before we can grant them autonomy in the world. One approach to doing so is to find ad-hoc rules for the setting at hand. However, historically, the AI community has significantly benefited from adopting methodologies that generalize across applications. The concept of expected utility maximization has played a key part in this. By itself, this concept falls short for the purpose of moral decision making. In this paper, we have consid-

ered two (potentially complementary) paradigms for designing general moral decision making methodologies: extending game-theoretic solution concepts to incorporate ethical aspects, and using machine learning on human-labeled instances. Much work remains to be done on both of these, and still other paradigms may exist. All the same, these two paradigms show promise for designing moral AI.

Acknowledgments

This work is partially supported by the project "How to Build Ethics into Robust Artificial Intelligence" funded by the Future of Life Institute. Conitzer is also thankful for support from ARO under grants W911NF-12-1-0550 and W911NF-11-1-0332, NSF under awards IIS-1527434 and CCF-1337215, and a Guggenheim Fellowship.

References

Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–26, 2007.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142, 1995.

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

Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. In W. Ramsey and K. Frankish, editors, *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2014.

Felix Brandt, Vincent Conitzer, and Ulle Endriss. Computational social choice. In Gerhard Weiss, editor, *Multiagent Systems*, pages 213–283. MIT Press, 2013.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, 2015.

Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

Swarat Chaudhuri and Moshe Vardi. Reasoning about machine ethics, 2014. In *Principles of Programming Languages (POPL) - Off the Beaten Track (OBT)*.

Scott Clifford, Vijeth Iyengar, Roberto E. Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, 2015. Available online at http://link.springer.com/article/10.3758/s13428-014-0551-2.

John P. Dickerson and Tuomas Sandholm. FutureMatch: Combining human value judgments and machine learning to match in dynamic environments. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 622–628, Austin, TX, USA, 2015.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, October 1991.

Bernard Gert. *Common Morality: Deciding What to Do*. Oxford University Press, 2004.

Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian C. Williams. Embedding ethical principles in collective decision support systems. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4147–4151, Phoenix, AZ, USA, 2016.

Marcello Guarini. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21(4):22–28, 2006.

Jonathan Haidt and Craig Joseph. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–56, 2004.

Daniel Kahneman and Amos Tversky. *Choices, Values, and Frames*. Cambridge University Press, 2000.

Joshua Letchford, Vincent Conitzer, and Kamal Jain. An ethical game-theoretic solution concept for two-player perfect-information games. In *Proceedings of the Fourth Workshop on Internet and Network Economics (WINE)*, pages 696–707, Shanghai, China, 2008.

John Mikhail. Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, 11(4):143–152, 2007.

Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

James H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

Francesca Rossi. Moral preferences. In *The 10th Workshop on Advances in Preference Handling (MPREF)*, New York, NY, USA, 2016. Available online at http://www.mpref-2016.preflib.org/wp-content/uploads/2016/06/paper-15.pdf.

Walter Sinnott-Armstrong. *Moral Dilemmas*. Basil Blackwell, 1988.

Walter Sinnott-Armstrong. Framing moral intuitions. In W. Sinnott-Armstrong, editor, *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pages 47–76. MIT Press, 2008.

Oliviero Stock, Marco Guerini, and Fabio Pianesi. Ethical dilemmas for adaptive persuasion systems. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4157–4161, Phoenix, AZ, USA, 2016.

Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.