Using Human Subjects' Judgments for Automated Moral Decision Making

Vincent Conitzer
Duke University

Jana Schaich Borg
Duke University

Walter Sinnott-Armstrong
Duke University

In many domains where artificial intelligence is being deployed, it is clear that the pursuit of simplistic objectives results in outcomes that are undesirable from a societal point of view. In machine learning, minimizing overall error rates may result in technologies working well for the majority of the population but poorly for minorities, putting the latter at a disadvantage. For example, speech recognition software may work poorly for those with nonstandard dialects or accents. Automated pricing to match supply and demand may result in price gouging. For example, when many people simultaneously want to leave an area due to a terrorist attack, an algorithm may drastically increase the price of rides out of the area to reduce demand and increase supply (see an article here). Such outcomes are often seen as morally wrong and unfair.

In some domains, it is possible to add patches to the system to address the most egregious cases of injustice or unfairness. For example, in the case of a recognized terrorist attack, a ride-hailing service could simply subsidize rides from the area and thereby get some good PR at relatively low cost. Such patches may be desirable, but they may also obscure the deeper problem that we have not figured out what we really want the algorithm to optimize for, leaving other problematic cases unaddressed. For example, what if the subway system suffers from unexpected flooding and many more people need rides? Is it wrong for the algorithm to drastically raise prices in this case? If not, what makes it different from the terrorist example? If it is wrong, should the ride-hailing company really be the one subsidizing rides, or should it be the city? Do we need more details about the case, for example how many people would be left in areas that are unsafe at night?

These thought experiments quickly make it clear that no simple objective function may capture exactly what we would like to see happen. Moreover, not everyone will agree what the right thing to do is. Transferring insights from one domain (ride-hailing) to another (speech recognition) is even more daunting. While we may eventually develop an exact, operationalizable, all-encompassing theory of ethics that everyone agrees on, those who design the algorithms running important parts of our lives need solutions sooner than that.

One methodology that can be quite generally applied is to present human subjects with a number of scenarios in the domain of interest, and ask them what the right thing to do is in each case. The "Moral Machine" website developed at MIT (`http://moralmachine.mit.edu/`) is a clear example of this (with some caveats, discussed below). Here, a subject is invited to imagine a self-driving car that has somehow arrived in a situation where it needs to take one of two undesirable courses of action—e.g., plow through a number of pedestrians crossing the road, killing them, or crash into a concrete barrier, killing the vehicle's occupants. Details of the scenario are randomly varied. (How many pedestrians? Which gender? How old are they? Did they cross the road illegally? Etc.)

Interesting insights can be obtained from these responses [1]. But also, given the responses, we can use machine learning techniques to model what drives each subject's decisions, and predict the decisions she would make in other scenarios (in the same domain). This model of the subject could then be used to guide the decisions of an AI system (say, a self-driving car, or a pricing algorithm) in the real world. While this may be an appealing outline of an approach, and one that can draw on a rich literature in preference elicitation [7], it requires many technical aspects to be addressed.

- We generally do not want to base morally-laden decisions by an AI system on a single subject's preferences. On the other hand, if we have accurate models of multiple subjects' preferences, they may disagree with each other on what the right thing to do is in a given scenario. Nonetheless, our AI system needs to make a single choice. One possible solution is to let the models of multiple subjects *vote* over the possible choices [3]. But exactly how should this be done? Whose preferences should count and what should be the voting rule used? How do we remove bias, prejudice, and confusion from the subjects' judgments? These are novel problems in computational social choice [2]. (A recent draft paper already explores some of these aspects in the context of the Moral Machine data above [5], as do we in our work discussed below.)

- From a machine learning perspective, what is the right framework? For example, how should we sample subjects? Also, suppose we can provide more than a point prediction of what a given individual would consider the right course of action—perhaps we have a distribution, where we predict there is a 60% chance the person would prefer option A. Should this then count as only 60% of a vote for A? If so would this create incentives for the person to answer differently, in order to have her opinion counted more?

- We need to ensure that the decisions faced by human subjects can actually inform the algorithm's decisions. Self-driving cars are unlikely to face such black-and-white choices as are presented on the Moral Machine website: there would likely be significant uncertainty about the outcomes of each action. This, by itself, does not mean the data is not useful for the purpose. The algorithm may be better able to deal with the probabilistic-reasoning aspects of the problem than most human subjects, and just need guidance on the relative valuations of outcomes. If so, the best scenarios to present to human subjects are not necessarily representative of actual scenarios the AI system will encounter in practice. On the other hand, we may wish to intentionally hide features that we feel should be irrelevant, such as gender, from subjects. We lack a good general theory of how to optimally design scenarios presented to subjects to guide AI systems in practice.

It is not hard to see that all these aspects interact. We have recently done a proof-of-concept case study [4] where we try to address all these issues end-to-end. We did this in the context of *kidney exchanges* [6]. Kidney exchanges are designed for the following situation. Suppose there is a patient in need of a kidney transplant who has a willing live donor, but due to medical incompatibility (e.g., blood types) the donor is not able to give to the patient. It is possible that the patient can swap donors with another patient in the same situation, so that both transplants are now medically possible. More complex arrangements to match patients to donors are also possible. Given everyone's medical data, the problem of finding an optimal matching is computationally hard, and algorithms from the AI community are already being used to address this problem.

... but what is an optimal matching? The first criterion one might think of is to simply maximize the number of transplants. But many people would be willing to prioritize a young, highly productive, otherwise healthy patient with dependents over an old patient with a variety of other serious health issues, some self-inflicted due to heavy drinking and smoking, who is unemployed, has no dependents, and a significant criminal record. Perhaps they would at least break a tie in favor of the former. If so, which are the attributes that we are comfortable with contributing to this decision, and which attributes should be ignored? We asked Mechanical Turk subjects this question to arrive at a list of attributes that might be taken into account. Then, we asked a different set of MTurk subjects to choose between imaginary patients that differed only in these approved attributes. We subsequently translated these decisions into weights on patients that the matching algorithm could use, for tiebreaking purposes only, and analyzed the effects on the exchange as a whole. We found that patient-donor pairs with certain blood type combinations were significantly affected (for better or worse) by this prioritization, whereas those with other blood type combinations were mostly unaffected. While the exercise was instructive and suggested the approach would in principle be feasible, it also revealed many nontrivial aspects of the general approach that future research should address.

References

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

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

[3] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4831–4835, San Francisco, CA, USA, 2017. Blue Sky track.

[4] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. 2017. In submission.

[5] Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan D'Souza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. 2017. Working paper, downloaded from http://procaccia.info/papers/ethics.pdf, October 9, 2017.

[6] Alvin E. Roth, Tayfun Sonmez, and M. Utku Ünver. Kidney exchange. *Quarterly Journal of Economics*, 119(2):457–488, 2004.

[7] Tuomas Sandholm and Craig Boutilier. Preference elicitation in combinatorial auctions. In Peter Cramton, Yoav Shoham, and Richard Steinberg, editors, *Combinatorial Auctions*, chapter 10, pages 233–263. MIT Press, 2006.