Moral Autonomy and Equality of Opportunity for Algorithms in Autonomous Vehicles

Martim BRANDÃO[a,b,1]
[a] *Oxford Robotics Institute, University of Oxford, United Kingdom*
[b] *Research Institute for Science and Engineering, Waseda University, Japan*

Abstract. This paper addresses two issues with the development of ethical algorithms for autonomous vehicles. One is that of uncertainty in the choice of ethical theories and utility functions. Using notions of moral diversity, normative uncertainty, and autonomy, we argue that each vehicle user should be allowed to choose the ethical views by which the vehicle should act. We then deal with the issue of indirect discrimination in ethical algorithms. Here we argue that equality of opportunity is a helpful concept, which could be applied as an algorithm constraint to avoid discrimination on protected characteristics.

Keywords. Autonomous vehicles, crash optimization, trolley problem, normative uncertainty, moral autonomy, indirect discrimination, equality of opportunity

1. Introduction

There has recently been an active discussion on the ethical issues of autonomous vehicles (AVs) [1, 2]. These vehicles will likely face moral dilemmas similar to trolley problems with extra considerations of risk, situational features and responsibility [3]. In particular, they might occasionally face the unavoidable option between crashing against one of several (groups of) people with different personal characteristics, or of transferring risk from pedestrians to passengers and vice-versa. In such cases, it is important to come up with morally relevant factors and ethical decision rules that reduce harm in a just way. The task is challenging because these decisions will need to deal with factors of chance, wellbeing, fairness and rights.

While we can try to avoid the occurrence of moral dilemmas through changes in design and infrastructure [4], it is unlikely that any infrastructure will be completely dilemma-free. For example, even if cars are only driven autonomously in dedicated roads, dilemmas might occur when one car needs to stop due to a malfunction and other cars do not have enough time to respond safely. Or, when a person or animal unexpectedly crosses an AV-dedicated road, regardless of whether such was permitted. It is also important to point out that not programming AVs with explicit decision rules to such dilemmas will still be making a moral stance depending on the distribution of the crash outcomes. For example, it is likely that dilemma-blind cars would not decide

[1] Martim Brandão, Oxford Robotics Institute, The George Building, 23 Banbury Road, Oxford, OX2 6NN, United Kingdom. E-mail: martim@robots.ox.ac.uk

to purposely hit a wall in order to save a life, as that would increase the risk of harm to passengers and the car. This would implicitly give preference to save the lives of passengers, which indirectly correlates with socio-economic characteristics of people likely to ride AVs. Another problem with not programming such decision rules is related to moral uncertainty and the possibly high cost of not doing so [5].

Our goal in this paper is to discuss two problems that arise when choosing decision rules to trolley problems in AVs: moral uncertainty (Section 2) and indirect discrimination (Section 3).

Moral uncertainty is related to the fact that we as car users, programmers, legislators or society in general, may not be sure of how to value and compare different actions and outcomes. It is also related to the concept of normative uncertainty [6], which is the uncertainty between ethical frameworks used to distribute risk given a certain metric (of harm, utility, wellbeing, etc.). In Section 2 we will argue that under this uncertainty, it would be morally questionable to enforce a decision rule on cars and thus violate the autonomy of car users[2]. We also discuss how giving autonomy to car users, within certain bounds, would be justified within virtue ethics.

Indirect discrimination is the phenomenon where outcomes of crashes correlate with personal characteristics of the people involved, even if those were not used in the decision process. We will introduce some examples of these in Section 3. We will also argue for the need of protecting certain characteristics and discuss how these could be implemented.

Throughout the paper we will make several comparisons between the moral decision problems in autonomous vehicles and those made in healthcare. On the one hand both deal with the distribution of wellbeing. On the other, in both there is enough time before the dilemmas take place for institutions and moral agents (car-users, physicians) to think about which factors to consider and equalize when setting a decision rule. Additionally, there are lessons to be learned for policy and governance of moral algorithms, regarding the dynamic and local nature of agreement on decision rules, and the discussion of persons' rights and moral autonomy.

2. Arguing for Morally Autonomous AV Users

2.1. The Uncertainty problem

While it is understood that autonomous vehicles should avoid negatively affecting people's wellbeing, defining wellbeing is not a simple task [7]. Wellbeing could be for example the probability of survival on the crash, as in [8], but there are many other justifiable measures of wellbeing. Here it is worth to look at the literature on ethics of healthcare resource allocation [7, 9], where different definitions have been argued for: number of life-years saved, quality-adjusted-life-years (QALYs), final health state, number of people with a minimal age, etc. In autonomous vehicles, the problem can be

[2] A similar argument is made in [16] with a focus on hybrid man-machine systems where the machine lacks moral decision-making capabilities. Note that in this paper we focus on moral autonomy regardless of machine capability. In other words, we argue that moral preferences of car users or all people involved in an accident should always be considered because of that uncertainty, while [16] implies that an AV with sufficient capabilities would have enough moral authority to force specific rules regardless of the preferences of the people involved. In this paper we also discuss objections and a virtue ethics argument for personalization which [16] does not consider.

further complicated if one includes animal rights into the equation, where it becomes necessary to balance the urgency of saving humans and other animals, or even artifacts such as monuments with high historical or ideological value. To keep our definition general, we will from now on refer to "utility functions" [8] instead of wellbeing so that different factors can be considered.

Not only are there several ways to define the morally relevant factors and to represent them using "utility functions", there is also no agreement on which one is right, or whether there could be a single acceptable function. In healthcare, different states adopt different functions and these are frequently revised [9]. To complicate the issue further, different ethical theories exist to make decisions based on a utility function: maximizing the sum [10], the average, a telic-egalitarian sum [11, 12] (utility plus a penalty on its inequality over the population), a minimax [8, 13], etc. Not only does each decision rule have different ethical implications, but arguably for parametric rules (e.g. utilitarian weighted sums and prioritarianism) there is an infinite number of parameter values where each makes a different ethical assumption.

Recent work in the community working on AVs also points in different directions. Some authors have crowdsourced trolley problem decisions with different numbers of people and personal characteristics to come up with a single utility function [14]. The approach is essentially the same as QALY's, although narrower in scope: to come up with an aggregated decision rule for distributing wellbeing by the use of questionnaires. Other work in the field applies a Rawlsian minimax decision rule to the probability of survival [8], and random decision rules have also been discussed [15].

The question of defining utility functions and decision rules is then complicated by the fact there might be several inter-related utility functions and several (if not infinite) possible decision rules based on those. Our assumption in the following arguments is that there will be several morally acceptable positions: whether because a single "true" ethical algorithm does not exist, or because there is uncertainty on which ethical theory we should choose. The latter idea, of normative uncertainty, has also been recently explored in the philosophy literature [6]. Based on it, the authors proposed using confidence values on ethical frameworks to guide decision-making. Again, there could be several acceptable confidence values for each ethical theory.

Next, we will consider two arguments for allowing car users to choose which decision rules (i.e. utility functions and ethical theories) to adopt. This could be implemented in practice by different algorithms being offered by different car providers, or personalized "car settings" available to the driver as previously proposed in [16]. Of course, we are not saying that in autonomous vehicles anything goes, but only that there might be a constrained set of morally acceptable decision rules in this context. We will later discuss how such personalized algorithms could be extended to account for preferences of all people involved in an accident.

2.2. The Autonomy Argument

We consider the following autonomy argument for allowing a diversity of moral choice to car users:

> If several acceptable decision rules exist, then it would be a violation of car users' moral autonomy to force a particular option on all users. Since there is no choice that is worthier than others, then users should be allowed to choose the ethical decision rule adopted by the car within the acceptable set.

Here we are echoing the argument used in healthcare towards a diversity of moral choice [7]. The same argument is used there, that since several morally tolerable positions exist for healthcare cost allocation and treatment decisions, then it would be a violation of the moral agent's autonomy to force one particular option such as the majority's preference or an institution's preference. Furthermore, justifying that violation is problematic since it is not clear whether decision rules of different people should be aggregated together into a single rule by any morally relevant reason. Note that this is in strike contrast with some of the recent work in autonomous vehicles, such as crowdsourced ethical decisions in simulated scenarios [14]. If several morally justifiable choices in trolley problems exist to begin with (since you allow people to vote for them) then the "aggregation" should happen naturally by each agent acting according to his selected morally acceptable principles, rather than subjecting each agent to act according to an aggregated policy.

So according to this argument it should be up to the moral agent to decide which utility function and decision rule he or she will choose from a list of acceptable ones, or to rank or weight among the several options. Further still, different car providers could adopt different rules, as long as all acceptable positions were available in the market (and as long as they all obeyed traffic rules and the law). Regardless of the implementation, we think that allowing for diversity of moral choice might attenuate other problems found in autonomous vehicles. One is the problem of dealing with cultural variation inside and between countries, and the other is that of attribution of responsibility for accidents since the role of each party is clear[3]. The state or policy maker is responsible for defining the right set of morally relevant factors, utility functions and/or decision rules, the manufacturer is responsible for correct implementation, and the user to decide his ethical position or degree of uncertainty over ethical principles. The user will always have done a morally acceptable choice unless the choices given by the state were wrong to begin with, while the car-maker will be accountable in case the outcomes do not respect the policy.

2.3. The Virtue Ethics Argument

Allowing moral diversity in autonomous vehicles is also consistent with virtue ethics, in that morality is something that is learned by doing, reflecting and discussing [17]. Note that we use "moral diversity" to mean a diversity of available algorithm options in AVs which (hence morally autonomous) AV users could opt from. A possible argument for diversity and autonomy could in this case be:

> "Moral virtue comes about as a result of habit" [17]. For car users to become virtuous, they should be able to try different decision rules and perfection them.

We can again compare the AV users' case with that of doctors. In states where doctors and first-responders are allowed to make certain life-death decisions [7], perhaps normal citizens should also be allowed such autonomy. Another argument from virtue ethics is based on increased political activity:

[3] For a discussion of responsibility in the context of hybrid man-machine systems, see [16].

> Political activity is an essential part of the good life [18]. Allowing a diversity of moral choice to car users would raise public discussion on ethics, justice and wellbeing. Therefore, allowing diversity would contribute to the good life.

Together, these two arguments say that through increased moral practice and discussion of morality, AV users and society as a whole could live better moral lives and potentially find better theories for moral choice.

Finally, we note that in this scenario of morally autonomous AV users, ethics experts and policy-makers still have a crucial role—of deciding the set of moral rules and factors over which there is uncertainty. This structure is similar to the view of politics which Aristotle [17] argues for, in that legislators are responsible for forming good habits in citizens and guide them to a virtuous life.

2.4. Objections

We now consider several objections to the idea of allowing AV users to opt between decision algorithms for moral dilemmas.

Ordinary people are not trained to think ethically. The objection assumes that only ethics experts or people with enough education in moral reasoning would be able to choose between such options correctly or provide acceptable justifications for their choices. One reply to such objection would be that since the set of options was already judged by experts to be morally acceptable, either because there is no agreement between experts or because of normative uncertainty and others, then there would be no option that was “more correct” than others—at least according to the knowledge available at the time the decision is made. Even for those AV users who do not want to reason about such issues, there could always be an option to follow the users’ or lawmakers’ majority. On the other hand, the introduction of these moral problems into the everyday lives of AV users could increase interest in the study of ethics or motivate the introduction of mandatory ethics education. Finally, from the point of view of virtue ethics and as we have said, the increased discussion and practice of moral problem solving would contribute to a virtuous life and eventually train people in ethics through practice.

The proposal breaks pedestrians’ autonomy. This is possibly the strongest objection to the proposal as we formulate it here, and it complains about the fact that AV users’ moral preferences and autonomy are respected while those of other people involved in a crash are not. We believe that no reasons related to fairness can be given for this, especially because personal characteristics of AV users might correlate with socio-economic factors that already favor them in large parts of society. We can only think of practical reasons for the proposal as we state it here (i.e. AV user autonomy), and those are technical feasibility and computation speed. However, imagining there were no technological limits, the autonomy proposal could be extended to all participants on a crash. For example, on the event of a crash, the moral preferences of all people involved in the crash could be aggregated in a way to find the best action to take according to everyone’s moral stance. Each person could hypothetically store his or her preferences (e.g. centrally or with them) on a system which was accessed by AVs during decision making. Assuming this was technically possible, it could be a fair procedure respecting everyone’s moral preferences as much as possible, at the cost of increased complexity. On the other hand, such an approach would have to deal with partiality and perverse incentives perhaps more strongly than on the AV-user case. This is because in principle the same moral stances should be available to usual-pedestrians and usual-AV-users, but in that case any agent would be able to game the decision-making system according to how often they expected to be driving or walking.

This proposal assumes there is a single AV user, which does not account for shared cars and public transport. The reply here could be a reframing of our previous point. Decision rules could either be aggregated (automatically voted for) according to all the users of a vehicle or of the city or state. Alternatively, the rules could be set by the transportation company, forcing the users to agree to them on use.

3. Arguing for Equality of Opportunity

3.1. The Problem of Indirect Discrimination

One problem that the previous section does not solve is that of indirect (or structural [19]) discrimination. This is the problem that people with certain personal characteristics might have higher probability of being victims of accidents with autonomous vehicles. Such indirect discrimination may happen just because of the unequal distribution of those characteristics within the population and the fact that they correlate with the outcomes of decisions used by vehicles. Consider the following examples:

1. Algorithm: maximize the expected number of lives saved. The algorithm might find ways to maximize this function by using visual appearance of people to predict their likelihood of surviving a collision, or their likelihood to have fast reflexes and evade the car. This could make people who look that way be more likely to have vehicles swerve towards them.
2. Algorithm: maximize the expected number of lives saved. The algorithm might lead vehicles to always swerve towards people who are alone, which could correlate with arbitrary psychological characteristics such as preference for walks alone.
3. Algorithm: maximize the expected number of life-years saved. The algorithm would prefer to swerve towards older people, or people who look older though they are not.
4. Algorithm: maximize the worse-case probability of survival (Rawlsian maximin). The algorithm would prefer to swerve towards bike drivers not wearing helmets than those wearing them [1]—the typical objection to maximin.
5. Algorithm: maximize the worse-case probability of survival (Rawlsian maximin). Even if statistics of crash survival are only slightly skewed or wrongly estimated, they will be correlated with some personal characteristics. For example, females could be predicted to have 0.0001 higher probability of survival than males of the same age, and therefore the algorithm would always prefer to swerve towards females.

Clearly such personal characteristics are morally arbitrary, which could pose a problem to the supposedly morally acceptable decision rules. Of course, it is of note that discrimination is not necessarily wrong per se, or wrong because it treats people arbitrarily, but that it is wrong only when it harms people with the characteristics in question [20]. But that is the case in our structural discrimination problem, since people with some characteristics will be harmed more in probability.

Allowing a diversity of moral choice in autonomous vehicles does not eliminate the indirect discrimination problem, because even if the same rule is not used by all vehicles, there will still be a moral position that is on average more in use. Therefore, personal characteristics related to these positions might be correlated with accident probability.

3.2. Equality of Opportunity on Protected Characteristics

Motivated by recent developments in the machine learning community, we argue that the notion of equality of opportunity as a constraint can be a solution to this problem. The idea is that whatever the utility function being optimized, the system could be made unbiased to protected personal characteristics by adding a fairness constraint [21, 22] or enforced normalization of predicted outcomes (e.g. enforce equal virtual outcomes of male and female pedestrians). Technically, the new (unbiased) decision rules would basically be solving an optimization problem on the original function subject to a constraint on the ex-ante (predicted) distribution of the protected characteristics. As an example, it could be decided that visual appearance is a protected characteristic, and thus that the decision probabilities should be the same whatever the visual appearance of the person (e.g. the distribution of pixel colors on a person seen through the robot's cameras).

Note that an important point in this solution is that algorithms should enforce ex-ante (predicted) fairness, not ex-post (based on past measurements of) fairness. The proper way to deal with uncertainty has been a subject of discussion in philosophy [23, 7, 24]. In our context, we mean that the decision should be corrected by predicting (e.g. through simulation) the outcome distribution of these characteristics over large samples "offline". This should not be done "online" by using recent statistics of accidents such as to equalize them as they happen. The latter, ex-post, approach would of course lead to morally arbitrary outcomes which would depend on the recent history of accidents—and are random events.

Besides correcting for unwanted discrimination, an advantage of such fairness-constrained methods is, in our opinion, their suitability to policy. The reason is that current machine learning-based algorithms would likely never use personal characteristics like age and gender directly and so a policy of forbidding its use would have no effect. Equality of opportunity constraints on protected characteristics would be a general enough framework to account for modern systems, which "learn" how to improve performance in terms of a given utility function from raw data without being given specific features. Finally, note that this method could be used together with moral diversity in autonomous vehicles, since each decision rule available to car users could be equalized to avoid biases in protected characteristics.

3.3. Alternatives

This proposal is subject to the usual objections to equality of opportunity and egalitarianism, such as the leveling down objection and libertarian objections. However, we can think of few alternatives to solving the indirect discrimination problem, except for affirmative-action-based constraints [22] or random actions [15]. Again, random choices in moral problems have been the subject of extensive discussion in the context of healthcare [7]. “Throwing the dice” to choose whose life to save is only justifiable when there are no morally relevant distinctions between the people involved, or when the uncertainty in these distinctions is too high [25]. This might well be the case in trolley problems, especially if all persons in a situation are not breaking any rules. The objections to the random approach are nevertheless serious, as is it can be interpreted as giving up on the search for a good moral rule or avoiding responsibility.

4. Conclusion

Disagreement between utilitarianism and its rivals has been going on for long, and it is questionable whether one ethical framework can be singled out until autonomous vehicles are deployed, or at all. The same problem pervades the choice of numeric representations of wellbeing or utility. We argue for responsible vehicle users who (have their vehicles) act according to their ethical views as long as they are justifiable. We see this as a better alternative than imposing a single decision rule and notion of utility on all vehicles, whether chosen by an institution or by aggregating preferences over the population. Such imposition would be both difficult to justify and violate moral agents’ autonomy. However, it is clear that ex-ante unfairness through indirect discrimination is a concern in autonomous vehicles. We believe recent work in enforcing constraints on equality of opportunity for protected characteristics (e.g. race, gender, visual appearance) in machine learning could be applied to solve this problem.

Finally, we note that our discussion of autonomy and equality is generally applicable to other value-laden technologies. For example, survey and rescue robots for disaster response probably need to make such life-death decisions by design, and indirect discrimination is also present due to correlations between geographical locations and socio-economic characteristics. Such examples will only increase as automation is applied across different areas of life, whether with explicit or implicit distributive impact.

References

[1] N.J. Goodall, Machine ethics and automated vehicles, in *Road vehicle automation*, Springer, 2014, 93–102.

[2] P. Lin, Why Ethics Matters for Autonomous Cars, in M. Maurer, J.C. Gerdes, B. Lenz and H. Winner (eds), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, 69–85.

[3] S. Nyholm and J. Smids, The ethics of accident-algorithms for self-driving cars: an applied trolley problem?, *Ethical Theory and Moral Practice* **19**(5) (2016), 1275–1289.

[4] V. Dignum, Responsible autonomy, in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, AAAI Press, 2017, 4698–4704.

[5] V. Bhargava and T.W. Kim, Autonomous vehicles and moral uncertainty, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (2017).

[6] W. MacAskill, *Normative Uncertainty*, PhD thesis, University of Oxford, 2014.

[7] W. Loh and J. Loh, Autonomy and responsibility in hybrid systems, in P. Lin, R. Jenkins, K. Abney (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (2017), 35-51.

[8] E. Elhauge, Allocating health care morally, *Cal L. Rev.* **82** (1994), 1449–1544.

[9] D. Leben, A Rawlsian algorithm for autonomous vehicles, *Ethics and Information Technology* **19**(2) (2017), 107–115.

[10] G. Bognar and I. Hirose, *The Ethics of Health Care Rationing: An Introduction*, Routledge, New York, 2014.

[11] M.J. Stuart, *Utilitarianism*, World Pub. Co., Cleveland 1863.

[12] D. Parfit, *Equality or priority?*, University of Kansas, 1995.

[13] I. Hirose, *Moral Aggregation*, Oxford University Press, Oxford, 2014.

[14] J. Rawls, *A Theory of Justice*, Harvard University Press, Harvard, 1971.

[15] R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar and A.D. Procaccia, A voting-based system for ethical decision making, *arXiv preprint arXiv:1709.06692* (2017).

[16] P. Lin, The robot car of tomorrow may just be programmed to hit you, *Wired Magazine* (2014). http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/.

[17] D. Ross, Aristotle: The Nicomachean Ethics, *Philosophy* **31**(116) (1956), 77–77.

[18] B. Jowett et al., *The politics of Aristotle*, Vol. 1, Clarendon, Oxford, 1885.

[19] H.-Y. Liu, Structural discrimination and autonomous vehicles: immunity devices, trump cards and crash optimisation, in J. Seibt, M. Nørskov, S. Schack Andersen, *What Social Robots Can and Should Do,* IOS Press, Amsterdam, 2016, 164–173.

[20] K. Lippert-Rasmussen, Discrimination: What is it and what makes it morally wrong?, in J. Ryberg, T.S. Petersen and C. Wolf (eds), *New Waves in Applied Ethics*, Palgrave-Macmillan, 2007.

[21] M. Hardt, E. Price, N. Srebro et al., Equality of opportunity in supervised learning, in M. Jordan, Y. LeCun, S. Solla *Advances in Neural Information Processing Systems*, MIT Press, Cambridge MA, 2016, 3315–3323.

[22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, Fairness through awareness, in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ACM, 2012, 214–226.

[23] J. Broome, Uncertainty and fairness, *The Economic Journal* **94**(375) (1984), 624–632.

[24] M. Hayenhjelm, What is a fair distribution of risk?, in S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (eds), *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, Springer Netherlands, Dordrecht, 2012, 909–929.

[25] J. Broome, Selecting people randomly, *Ethics* **95**(1) (1984), 38–55.