Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Understanding Autonomous Driving as Institutional Activity: Opening New Ways to React to Discriminatory Concerns in Autonomous Driving

Manuel Dietrich

2020

Preprint:

This is an accepted article published in Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020 . The final authenticated version is available online at: https://doi.org/10.3233/FAIA200934

Culturally Sustainable Social Robotics
M. Nørskov et al. (Eds.)
IOS Press, 2020

© 2020 The authors and IOS Press. All rights reserved.
doi:10.3233/FAIA200934

Understanding Autonomous Driving as Institutional Activity: Opening New Ways to React to Discriminatory Concerns in Autonomous Driving

Manuel DIETRICH[a,1]
[a] *Honda Research Institute Europe, Germany*

Abstract. We will elaborate on why it is appropriate to be concerned about structural discrimination induced by future autonomous vehicles—not only in the context of crash-optimization. In order to understand discrimination effects and potential strategies to reduce them, a novel conceptual perspective on how to frame autonomous driving is introduced. According to it, autonomous driving is considered an institutional activity. This perspective enables us to connect to theories about institutional justice—and open new ways to react to discriminatory concerns. To illustrate this direction, we sketch how to apply institutional justice to an autonomous highway driving system.

Keywords. Autonomous driving, machine ethics, self-driving cars, autonomous vehicles, discrimination, biases, institution, Rawls, distributive justice

1. Introduction: Discrimination and Autonomous Driving

Since machine learning methods or artificial intelligence have become a functional part of technical systems and applications, people have been concerned that their outcomes or services are biased or even discriminating [1, 2]. Systems that decide or behave autonomously, are under special scrutiny if they favor certain groups of users or otherwise affected people over others.

This discriminatory potential becomes especially apparent when artificial intelligence methods treat people of certain races or ethnic groups differently. An illustrative example is an intelligent system that was used for judging beauty and was biased against darker-skinned contestants [2: 2].[2] Discriminatory concerns have also been raised when using artificial intelligence for credit ranking [3] or criminal profiling.[3]

Friedman and Nissenbaum were among the first to attend to possible biases of computing systems and suggested the following general definition:

---

[1] Corresponding author: Manuel Dietrich, Honda Research Institute Europe, Carl-Legien-Straße 30, 63073 Offenbach am Main, Germany; e-mail: manuel.dietrich@honda-ri.de.

[2] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

[3] See here: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. The propublica article was strongly criticized by researchers because of methodical errors [4].

> A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate [5: 332].

With reference to the above examples, an undesirable outcome would be a rejected loan application or the involvement in criminal investigations as suspect. It is considered as unreasonable or inappropriate if the undesirable outcome is not based on relevant criteria for the application. For instance, an individual's loan repayment history could be relevant for a credit decision, whereas the membership in a group or a certain personal attribute, which has nothing to do with the decision, would not. The exact criteria will heavily dependent on the application. There are defined attributes, for instance race, gender, or religion, which should not have an impact on any decision, otherwise the decision is legally considered as discriminating (see: [6: 1569]).

One reason for biased autonomous systems is the usage of "unbalanced data" for training the models—that means to utilize training data in which certain groups are under- or overrepresented [2: 7]. An example is a face recognition technology, which only performs reliably with Caucasian skins, because the data for training contained a much higher proportion of faces with that skin type [7]. Another reason for algorithmic biases is that systems are based on data of human influenced processes and simply "maintain some of the same biases that permeate society" [8: 1].

In the context of autonomous driving, unbalanced training data can increase the probability that people with dark skin tones are overlooked by the vehicle. For instance, Wilson et al. uncovered biases in pedestrian detection. They found out that state-of-the-art detection systems perform better with white people and struggle to identify pedestrians with darker skin tones [9].

In this paper we focus on a specific type of discrimination in autonomous driving which was put forward by Hin-Yan Liu [10, 11]. He raised the concern that autonomous vehicles will discriminate in a structural way, induced by the fact that future autonomous vehicles will not appear on our roads as individual entities, but rather as a fleet of vehicles which follow similar behavioral principles. Liu summarizes his concern as follows:

> Structural discrimination arising from the coordinated modes of autonomous vehicle behavior […]. This leads to the potential for hitherto individuated outcomes to be networked and thereby multiplied across fleets of vehicles. The aggregated effects of such algorithmic policy preferences will thus cumulate in the reallocation of benefits and burdens to certain categories of persons in a relatively stable manner [11: 150].

The goal of this paper is to further discuss this type of structural discrimination by looking at it from a new conceptual perspective. This novel view frames autonomous driving as an institutional activity. The connection to institutions and its operating principles allows us to react to discriminatory concerns from an institutional justice and fairness perspective. In order to show the potentials of utilizing principles of justice, we will illustrate how to apply such principles to a driver assistant system for partially autonomous highway driving and discuss the results.

2. Autonomous Driving as Institutional Activity

Understanding institutions and institutional activity is a research topic in sociology and social science [12, 13] as well as in philosophy. In philosophy, institutional activity is discussed in the tradition of action theory and considered as a form of collective activity [14]. In this section, we will look at institutions from a sociological perspective, which means to consider them mostly as social institutions.

Social institutions play a prominent role in John Rawls' theory on justice [15]. The principles of justice he developed are in particular relevant to social institutions; this means that the principles should guide their operations, for instance the distribution of social goods. Surprisingly Rawls did not provide a deeper theoretical definition of what social institutions are in his main work. However, in some of his examples he gets more concrete and for example refers to health, education, and training as primary goods, (i.e. 15: 54]) which should be fairly distributed in society. It is straightforward that institutions which are concerned with providing these goods, for instance educational institutions like universities or schools, are major social institutions.

Even though institutions are a key element in many theories which explain the basic structure of society, there is no shared understanding on the defining elements of an institution [16]. However, we can extract some features which are brought up by many authors. Especially two decisive elements are of relevance: Institutions (social) are constituted by a set of behavioral rules, either explicit policies (formal institutions) or implicit social practices, for instance an "institutional culture" [16: 37]. Additionally, they are constituted by the individuals who occupy roles in the collective of the institution and who, as "proxy agents" ([17], discussed e.g. in Chap. 13), apply or follow the behavioral rules (norms and policies) and so form the institutional activity. Institutional roles are for instance teachers, lecturers, doctors, civil servants or judges. It is important to emphasize that "institution[s] provide a framework within which individuals act" but "they do not fully determine the actions of individuals" [16: 38]. Individuals are the executive bodies, but they always act on behalf of the institution in order to fulfil their roles.

This structure of roles and rules makes the activity of an institution persistent and predictable and a stabilizing element for society (see e.g. [18: 2]). As central stable entities, institutions have a strong impact on people's lives and the way how citizens live together in society. Because of this, rules and policies have to be chosen with great care—which makes institutional justice important.

In the next section we will discuss why this perspective is relevant in the autonomous driving context.

2.1. Why Should We Frame Autonomous Driving as Institutional Activity?

Based on the understanding of social institution, we will explain why we think it is reasonable to consider autonomous driving as an institutional activity and what follows from that.

We agree with Liu (see Sec. 1), who argues that we will not meet autonomous vehicles as individual entities, rather they will be one entity of a fleet of vehicles which all follow the same rules and policies in their driving decisions. It is simply a rational business choice of manufacturers to equip all their autonomous cars or at least a series of cars with the same, well optimized decision-making systems. We might also see cooperation between manufacturers and with that a more extensive aggregation. Another

possibility is that the car manufacturers themselves do not develop the modules which are central to allow self-driving functionalities, but rather that the main components (possibly including its policies) are provided by third-party supplier companies.

The constellation of the individual vehicles on one side and on the other side a general driving policy, matches the key characteristics of an institution we have discussed before. This is based on the premise that we consider autonomous vehicles as individual agents, which are certainly programmed to follow a behavior policy but are not determined by it in a strict sense. Liu argues, "programming of the autonomous vehicle only makes certain outcomes more likely than others but does not determine which outcome manifests" [11: 153]. Human agents are influenced by their personal history, ideological beliefs or character when fulfilling their instructional role. Civil servants who are handling shelter application have to follow strict policies, but the decisions are not independent of their personal judgement. Autonomous vehicles operate with a much lower degree of freedom. However, a system which contains algorithms formed through data-driven learning, for instance computer vision-based scene interpretation, will construe policies with a significant degree of variations when released to the real road.

We think it is reasonable to look at the social aspect of the described institutional practice. Let us recall the thoughts outlined before (see: Sec. 2 intro): social institutions are a key stabilizing element for societies, because they are organized to be persistent and take over functions which serve the public interest. Considered as social institutions are not only institutions providing service or goods but also institutions who have, by the means of their activities, impact on citizens' key values and rights. For instance, if institutional practices affect people's opportunities to be involved in society (social participation) or which share they receive of a communities' or countries' wealth (e.g. tax institutions).

Autonomous vehicles might not provide services which are essential for society or produce a social good—that means in a strict sense they do not serve the public interest. Nevertheless, their activities could crucially effect citizens' values and rights. This becomes apparent in situations in which an autonomous vehicle cannot avoid a crash and will have to decide how to crash and whom to harm [19].

The impact on people's values and rights caused by autonomous vehicles is however not only evident in crash situations. Many driving decisions entail aspects of significant impact. This is the case because most of the vehicle's behavior reshapes the operational space of other traffic participants and influences how they are able to fulfill their mobility needs and interests [20, 21]. We previously (see [20]) identified three key mobility or driving needs: First, safety, that means the probability to not get into a hazardous situation. Second, utility, the progress towards a certain destination, and third, the driving comfort. In most circumstances the individual's driving needs conflict with those of other road participants due to the limited road space. Many driving decisions are not regulated by law but have an impact on how others can fulfil their driving needs, for instance when it is appropriate to break or do a lane-change that allows others to proceed faster.

A fleet of vehicles that follows the same set of rules or policies, could multiply effects in a certain direction, maybe advantaging cars of a certain brand or certain other traffic participants and discriminating others. This is why we should treat autonomous driving as an institutional activity and more specifically frame it as an activity similar to those of social institutions.

In order to explain more what the institutional perspective is, we briefly discuss the difference between human driving and autonomous driving in this regard. Should we

also consider human driving, as it currently exists, as an institutional activity? Indeed, we can see that some key characteristics of an institutional activity are met: Human decisions are constituted or in other words regulated by traffic rules (law)—whereby drivers can decide to follow them or not, with all its consequences. The difference is that normal traffic participants are not part or even representatives of the institution, which the traffic rule system is constituting. This is different for traffic police officers, who enforce the traffic rules, with a degree of freedom of how to. Police officers indeed occupy an institutional role and so their practice of law enforcement, should be categorized as an institutional activity. Since human drivers are no role takers, or in the words used before, act on behalf of the institution, it is also not adequate to frame the multiplicity of individual driving decisions of humans as a collective or even institutional activity.

What follows from applying an institutional perspective? Seeing autonomous driving from an institutional perspective can be utilized to approach the discriminatory concerns. Social institutions have received special attention from moral and political philosophers. They developed theories about which rules and policies institutions should follow in order to fulfill their function in a just manner and produce fair outcomes. Being non-discriminating is one of the key principles for institutional justice.

Very well-known is John Rawls' fairness as justice approach for the activities of social institutions.[4] In the following quote, he explains how he sees the relationship between justice, society, and social institutions:

> For us the primary subject of justice is the basic structure of society, or more exactly, the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation [15: 7].

Rawls also emphasizes the difference between institutional and individual justice: "The principles of justice for institutions must not be confused with the principles which apply to individuals and their actions in particular circumstances. These two kinds of principles apply to different subjects and must be discussed separately" [15: 47]. To elaborate this difference more: An individual ethics perspective refers to individual beliefs about morality and right and wrong. So-called experimental ethics is one way to learn about beliefs with regard to situations of ethical relevance [22]. One influential approach on the ethics of self-driving cars, the so-called trolley case experiments, also known as the MIT moral machines experiments [23, 24], is such an experimental or individual ethics approach. In this series of experiments, study participants all over the world were asked to decide in a hypothetical situation who an autonomous vehicle might harm if they would come in a dilemma situation, a situation where all possible outcomes will lead to a fatality.[5] One could use the results of these experiments, to build the foundation of how future autonomous vehicles behave on the road. Yet, the results are often very culture dependent (which was shown for autonomous driving [24]), and more importantly, as Rawls said, ethical principles which apply for individuals are different to those of institutions. Following the institutional understanding of autonomous driving has the consequence that we should not draw too much attention to such experimental findings, rather we should look on how social institutions justify their policies.

By no longer assessing autonomous driving from an individual ethics but rather from an institutional perspective, other normative *tools*, which we think are more suitable,

[4] Rawls explicitly explains that his theory is concerned with the "justice of institutions" [15: 3].

[5] For a critical discussion on trolley-cases and autonomous driving, see: [25].

become available. There are many other authors, who, some of them in Rawlsian tradition, developed theories about institutional justice (To mention only three of the more recent prominent authors, see [26-28]). Understanding autonomous driving as institutional activity enables us to connect to well established theories of institutional justice to tackle discriminatory concerns. In the next chapter, we will sketch how to apply concepts of institutional justice to the autonomous driving context.

3. A Distributive Justice Approach to Reducing Certain Discriminatory Effects

In this practical example, we will describe a driving policy, which is based on Rawls' concept of distributive justice, which was originally developed in the context of social institutions. This applied fairness as distributive justice autonomous driving approach has been previously presented by us [20]. In this paper we look at it as an example for applying institutional ethics for autonomous driving. We were not the first who made use of Rawls' theory in the context of autonomous driving [30], however, by connecting autonomous driving and social institutions, the reasons why this is a qualified approach becomes even clearer. At the end we discuss if this approach is feasible to address some of the discriminatory concerns.

Before we continue, we will return to the discriminatory concerns related to autonomous driving described by Liu, presented at the beginning of this paper and discuss Liu's proposals of how to react to these concerns. Both ideas are about initializing policies for autonomous driving in a way that the vehicles will not be part of a collective activity—or how we have framed it, autonomous driving should not be an institutional activity.

- *Involving Customer's Choice:* The idea is that users of autonomous vehicles choose for themselves what rules or policies the vehicle should follow [11: 179]. This "Adjustable Ethics Settings" [31] approach could lead to the policies of every vehicle becoming very different. Yet this would only be the case if there are enough possibilities to choose from and the customers actually make use of the variety of options. Even when Liu argues that it is worth to think about such an approach, we think there are obvious downsides [11: 179]. It is hard for users to foresee what are the consequences of their choices. Human drivers might have a style and an implicit understanding of what they consider as right or wrong while driving, but this does not mean they can communicate them explicitly or even want that autonomous vehicles act similar to what they would do.
- *Instilling randomness:* Adding randomness to the behavior choices of autonomous cars is what Liu calls the lazy approach [11: 180]. It could help to avoid that all vehicles of a fleet act too similar. However, this would be hard to communicate to potential customers and the general public. Technologies that act unpredictably and consequently non-optimally, would violate the fundamental expectations people have about technical systems (technology as reliable element, see e.g. [32]).

Both directions have strong disadvantages, so de-institutionalizing autonomous driving might not to be a practicable way.

3.1. Applying Fairness as Distributive Justice for Autonomous Non-crash Driving

The system we use here as an example is an advanced adaptive cruise control system for partially autonomous driving on highways, which has been modified by a distributive justice policy. It is not an emergency or crash-optimization system which are often subject of an ethical reflection. However, it is suitable to illustrate our approach because, as discussed in Section 2.1., not only can the decision-making in crash situations pertain discrimination, rather autonomous driving policies which contain general rules can have discriminatory (cumulative) effects when multiplied through a fleet of vehicles. Autonomous vehicles could be discriminating when they exploit their technical supremacy ("privatization of public space", [11: 175]) and thereby hinder others in fulfilling their utility or comfort needs regarding mobility, thereby shifting or reshaping the opportunities of certain groups of citizens to have access to public space.

Next, we will briefly explain the fairness as justice autonomous driving architecture. The key idea is that a fair vehicle *respects* the fact that the public road is a social good. Thus, the fair vehicle makes driving decisions not only from the perspective of an autonomous vehicle but takes the driving needs of other traffic participants equally into account, whereby the predicted worst-off participant should profit the most from a driving decision (difference principle).

This idea of a fair vehicle derives from Rawls' two principles of a just society. Rawls' most important principle is that citizens have equal liberties and rights. For instance, one basic right of citizen is that they are treated equally in accordance with the rule of law [15: 266]. This means that a just society has to make sure that recipients are treated equally according to the rules that apply for them. With respect to fair autonomous driving, this is realized by estimating the driving opportunities of all traffic participants in the same way, that it is done for the autonomous vehicle itself. A decision is therefore taken by equally respecting the driving needs of all traffic participants.

In Rawls' theory the second subordinated principle for a just society is what he calls the difference principle. Applying the difference principle leads to an allocation of goods in which the most disadvantaged recipients, introduced by Rawls as the "worst-off" citizens in society, are supported the most [15]. Transferring the difference principle in the driving context leads to the following policy: The vehicle with the lowest expected driving prospects amongst all vehicles in the scene should be come off high as possible. In our approach, driving prospects are not attributed to certain traffic participants; rather they are calculated scene-by-scene, based on the driving need criteria including safety and utility, as will be elaborated in the next paragraph. In other words, our approach produces driving decisions in which the predicted worst-off vehicles (in every situation/scene an autonomous decision is made) profit the most from this decision.

3.1.1. System Architecture

The fair autonomous driving architecture is an extension of the intelligent cruise control system for highway driving.[6] To understand how the fairness approach has been applied, we will sketch how the original system performs. The system carries out three steps. First step is the Scene Detection, which means the autonomous vehicle senses its surroundings through computer vision and LIDAR and recognizes vehicles and other relevant scene elements. Second, is the Behavior Planning step, which optimizes up to three behavior

[6] The original system has been presented by Weisswange et al. [29] and the fairness extension by us [20].

options (one for straight driving, lane-change left and right respectively), that in turn balance costs and benefits for the autonomous vehicle. Third, is the Behavior Selection step, in which the car selects the option with lowest costs and communicates this option to the car's actuation system.

In order to apply the principle of fairness into the autonomous driving architecture described before, the Behavior Planning and the Selection steps have to be modified. The planning step is extended in order to involve other traffic participants. This means that the costs of the other traffic participants (driving opportunities) are predicted in the same way, making driving quality and safety needs comparable. The same original behavior planning module is used for this: a full behavior planning loop, that applies the same optimization mechanisms used for the autonomous vehicle planning. Based on the understanding about the other traffic participants' needs, the autonomous vehicle will select a driving maneuver considering all these needs. This mean a max-min rule is implemented, with which the autonomous vehicle always selects the option from which the vehicle with the lowest expected driving opportunities (highest costs) profits the most.

3.1.2. Example

An example shall illustrate this approach (Figure 1). The screenshot is taken by the front camera of the autonomous vehicle prototype. The situation can be described as follows: The autonomous vehicle is driving in the middle lane at a relatively high speed. The two other relevant vehicles in the scene are first a white car with a slower speed than the autonomous vehicle and second the truck in front of the white vehicle. The white vehicle is significantly faster than the truck.

The behavior planning of the autonomous vehicle results in two reasonable options, first a lane change maneuver to the left and second to continue the straight path. If considering only the own driving needs, the autonomous vehicle has higher costs when opting for a lane-change, because changing lane is always considered as less comfortable and less optimal from a utility perspective.

When including the other traffic participants' perspective, the assessment changes. If the autonomous vehicle would continue going straight, the white vehicle will have to break behind the truck, because it will not have the possibility to do a lane-change, since the autonomous vehicle blocks its way. Strong braking will, in the technical terminology, produce high costs. If the autonomous vehicle decides for the lane-change option, the white vehicle could perform the lane-change much earlier, a maneuver with less cost (less risk and higher utility) for the white vehicle. By letting the autonomous vehicle follow our fairness approach, the white vehicle's costs are included in its decision: The white vehicle is identified as the vehicle with the predicted maximal cost, it is the worst-off in the scene. Since the difference principle as our leading fairness policy is targeting for decisions where the worst-off profits the most, the autonomous vehicle will perform the lane-change to reduce the white vehicle's costs.



Figure 1. A screenshot taken by the front camera of the autonomous vehicle prototype.

4. Discussion and Conclusion

Finally, we discuss if such an approach can help to decrease discriminatory effects. To apply a decision rule which takes the needs of others equally into account, decides in favor of the worst-off, rather than solely targeting on self-optimization, can at least reduce some of the concerns, as we have sketched before.

What happens if the proposed fairness policy is applied to a fleet of vehicles? Based on how the system architecture and policy is designed to include the mobility needs of all the other participants, the fairness policy will prevent that certain vehicle types (e.g. with low engine power) or drivers (e.g. passive driver styles) might be systematically left behind. The discriminatory concern that a fleet of autonomous vehicles will intentionally or unintentionally exploit their technical supremacy, because of for instance better foresight capability compared to human-driven vehicles, can be refuted. This is ensured through a policy which targets a just distribution of road share, considering safety, utility and comfort factors and thus hinders autonomous vehicles to take hold of a higher proportion of the road.

We think our practical approach might reduce certain biases, but by no means all. Biases which result from the scene detection component (recognizing scene elements as described in 3.1), for example that certain vehicle types (e.g., motorcycles) are not detected as well as others, cannot be resolved at the moment. This is problematic, because the mobility needs of traffic participants who are not detected or detected later than others, will not be considered or considered too late by the autonomous vehicle, hindering it from taking the needs of all participants into account. For minimizing biases caused by computer vision through uncomplete or unbalanced data, other approaches and strategies are needed.

In our practical approach we did not consider driving scenarios with mixed traffic, for instance inner-city traffic, where traffic participants who are more vulnerable e.g. cyclists and pedestrians are involved. Including the perspective of others and considering their mobility needs is easier when traffic participants have similar capabilities and presumed similar needs in traffic. In asymmetric traffic, the exact parametrization of autonomous vehicle driving policies must be subjected to a careful consideration, because it is difficult to adequately respect factors or values that are important to other types of traffic participants. It is challenging to find a good and reasonable foundation to

weigh safety needs of, for instance, cyclists and cars sharing the same road, and to include this in the cost calculation.

However, looking at autonomous driving as an institutional activity allows to discuss behavioral policies from a social justice perspective. Considering the road as a shared social good, which should be fairly—i.e., here: equally—distributed between citizens, is one way of moving forward. We think the difference principle as distributive rule is, in the way we have used it for our concrete technical implementation, a reasonable choice.

References

[1] Holstein K, Wortman VJ, Daume H, Dudik M, Wallach H. Improving fairness in machine learning systems: what do industry practitioners need? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery; 2019. p. 1-16.
[2] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. In: Computing Research Repository, arXiv eprint: 1908.09635; 2019. p. 1-31.
[3] Chen J. Fair lending needs explainable models for responsible recommendation. arXiv eprint: 1809.04684; 2018. p. 1-4.
[4] Flores A, Bechtel K, Lowenkamp C. False positives, false negatives, and false analyses: a rejoinder to "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". Federal Probation Journal. 2016; 80(2):38-46.
[5] Friedman B, Nissenbaum H. Bias in computer systems. ACM Trans Inf Syst. 1996; 14(3):330-347.
[6] Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R. Fa*ir: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, New York: Association for Computing Machinery; 2017. p. 1569-78.
[7] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editor. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research 81. MLResearch Press; 2018. p. 1-15.
[8] Howard A, Borenstein J. The ugly truth about ourselves and our robotcreations: the problem of bias and social inequity. Sci Eng Ethics. 2018;24(5):1521-36.
[9] Wilson B, Hoffman J, Morgenstern J. Predictive inequity in object detection. arXiv eprint: 1902.11097; 2019. p. 1-13.
[10] Liu HY. Structural discrimination and autonomous vehicles: immunity devices, trump cards and crash optimisation. In: Seibt J, editor. Proceedings of Robophilosophy/TRANSOR 2016. Amsterdam: IOS Press. 2016. p. 164-73.
[11] Liu HY. Three types of structural discrimination introduced by autonomous vehicles. Uc Davis Law Review Online. 2018 May; 51:149-180.
[12] Durkheim E. Rules of sociological method. Contemporary social theory. New York: The Free Press; 1982.
[13] Weber M. The methodology of the social sciences. New York: The Free Press; 1949.
[14] Searle J. Collective intentions and actions. In: Cohen PR, Morgan J, Pollack M, editors. Intentions in communication. Cambridge: MIT Press; 1990. p. 401-15.
[15] Rawls J. A Theory of justice. Revised Edition. Cambridge: Harvard University Press; 1971.
[16] Miller S. Authority, discretion and accountability: the case of policing. In: Sampford C, Preston N, Bois C, editors, Public sector ethics: finding and implementing values. London: Routledge; 1998. p. 37-53.
[17] Ludwig K. From plural to institutional agency: collective action II. Oxford: Oxford University Press; 2017.
[18] Hechter M, Opp K, Wippler R, editors. Social institutions: their emergence, maintenance and effects. London: Taylor and Francis; 2018.
[19] Goodall N. Machine ethics and automated vehicles. In: Meyer G, Beiker S, editor. Road Vehicle Automation. Cham: Springer International Publishing; 2014. p. 93-10.
[20] Dietrich M, Weisswange TH. Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. Ethics and Information Technology. 2019; 21(3):227-39.
[21] Mladenovic MN, McPherson T. Engineering social justice into traffic control for self-driving vehicles? Science and Engineering Ethics. 2016; 22(4):1131-49.
[22] Lütge C, Rusch H, Uhl M, Luetge C, editors. Experimental ethics: toward an empirical moral philosophy. London: Palgrave Macmillan; 2014.

[23] Bonnefon JF, Shariff A, Rahwan, I. The social dilemma of autonomous vehicles. Science. 2016. 352(6293):1573-76.
[24] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I. The moral machine experiment. Nature. 2018. 563(7729):59-64.
[25] Nyholm S, Smids J. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? Ethical Theory and Moral Practice. 2016; 19(5):1275-89.
[26] Walzer M. Spheres of justice. New York: Basic Books; 1983.
[27] Nussbaum MC. Human functioning and social justice: In defense of aristotelian essentialism. Political Theory. 1992; 20(2):202-46.
[28] Margalit A. The decent society. Cambridge: Harvard University Press; 1998.
[29] Weisswange TH, Rebhan S, Bolder B, Steinhardt N, Joublin F, Schmuedderich J, Goerick C. Intelligent traffic flow assist: optimized highway driving using conditional behavior prediction. IEEE Intelligent Transportation Systems Magazine, early access. 2019.
[30] Leben D. A rawlsian algorithm for autonomous vehicles. Ethics and Information Technology. 2017. 19(2):107-15.
[31] Contissa G, Lagioia F, Sartor G. The ethical knob: ethically-customisable automated vehicles and the law. Artificial Intelligene and Law. 2017;25(3):365-378.
[32] Ihde D. Technology and the lifeworld: from garden to earth. Bloomington: Indiana University Press; 1990.