Reports and Communications

Social Science Computer Review
1-11

DOI: 10.1177/0894439320906508
journals.sagepub.com/home/ssc
SAGE

When AI Ethics Goes Astray: A Case Study of Autonomous Vehicles

Hubert Etienne[1]

Abstract

This article discusses the dangers of the Moral Machine (MM) experiment, alerting against both its uses for normative ends and the whole approach it is built upon to address ethical issues. It explores additional methodological limits of the experiment on top of those already identified by its authors; exhibits the dangers of computational moral systems for modern democracies, such as the "voting-based system" recently developed out of the MM's data; and provides reasons why ethical decision-making fundamentally excludes computational social choice methods.

Keywords

autonomous vehicle, self-driving car, AI ethics, Moral Machine, trolley problem

Autonomous vehicles (AVs)[1] represent a crucial challenge for artificial intelligence ethics, whose tremendous expected benefits justify the fierce competition between both manufacturers and national governments. While players have invested up to USD80 billion between August 2014 and June 2017 in the hope of conquering a solid share of a market forecasted to reach USD6.7 trillion by 2030 (Mohr et al., 2016, p. 6), governments have understood the necessity of adapting national regulations to support self-driving testing, foreseeing AV's high potential to contribute to economic growth and increase public safety. Such advances are, however, not without legal and ethical issues, whose most discussed so far has been that of moral responsibility and legal liability in the case of fatal accidents.

Anticipating complex situations in which AVs may not be able to avoid accidents and would consequently have to allocate harm between several groups of individuals, researchers have found in the "trolley problem" (Foot, 1967; Thomson, 1976, 1985), a theoretical framework to address the resulting moral dilemmas. Awad et al. (2018) expanded on this by developing the Moral Machine (MM), an online platform reproducing *trolley-style* thought experiments in various situations involving an AV, with the aim of establishing a global representation of moral preferences. The success of the experiment then became the starting point for

[1] Ecole Normale Supérieure, Paris, France

Corresponding Author:
Hubert Etienne, Ecole Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France.
Email: hubert.etienne@sciencespo.fr

Noothigattu et al. (2018) to develop a "voting-based system" (VBS), grounded in computational social choice theories, aiming to automate ethical decisions by aggregating the individuals' moral preferences collected by the MM.

This article presents a critique of both the MM and the VBS, highlighting their intrinsic limitations and revealing their deleterious effects on the debate. It brings to light the dangers proceeding from use of the MM data for normative purposes and the inner fallacy of attempting to automate ethical decision-making processes. The first part analyzes the construction of the AV moral dilemma and its current approach in the debate, from the advent of a moral imperative supporting the development of AVs to the conceptualization of the dilemma on the trolley problem's model and the deployment of the MM. The second part then criticizes the use of the MM data by the VBS and refutes the possibility of developing a legitimate computational system to automate ethical decisions. Finally, the last part denounces an instrumentalization of the ethical discourse, rejecting the "highest moral imperative" (HMI) underlying this project and exposing the distracting effects of the MM on public opinion, having both polarized the debate on erroneous principles and categories and prevented relevant ethical issues from receiving appropriate attention.

A Responsibility Issue Resulting From the Transfer of Autonomy

According to their proponents, AVs represent a unique opportunity to improve transportation by making traffic more fluid, inclusive, ecological, economical, and safer. Around 1.35 million people die every year in traffic accidents (World Health Organization, 2018, p. 4) while 94% of car accidents are said to result from human error (Singh, 2015, p. 1). Considering the numerous deaths that AVs may avoid, some of its partisans have even proclaimed a moral obligation to deploy them as soon as possible (Bonnefon et al., 2016, pp. 1575–1576; Shariff et al., 2017, p. 696), supporting relaxed regulations for manufacturers and lowering their liability in case of accident to avoid discouraging them (Hevelke & Nida-Rümelin, 2015, p. 629). Let us refer to this claim as the HMI, well illustrated by former National Highway Traffic Safety Administration chief regulator Mark Rosekin: "We can't stand idly by while we wait for the perfect [. . .] We lost 35,200 lives on our roads last year [. . .] How many lives might we be losing if we wait?" (M. Bauman & Youngblood, 2017). For some of the HMI proponents, the moral imperative applies as soon as AVs can reduce the net balance of total annual deaths by one, what leads them to develop simulation instruments designed to help policy makers identify this critical moment (Kalra & Groves, 2017), while others assert, on the same ground, that regular cars should be prohibited as soon as AVs become safer (Sparrow & Howard, 2017, pp. 209–210).

Although AV partisans advance a reasonable argument in favor of early deployment, some advocate that additional issues should first be considered before bringing AVs to market. AVs may indeed help avoid the majority of today's accidents resulting from human factors. However, they will not prevent all accidents, which could still occur through technical outages or meteorological conditions, sometimes resulting in complex situations where no evident choice could be unanimously preferred. One of them is analyzed by Lin (2015) in a thought experiment where a subject is driving a nonautonomous vehicle (NAV) on the central lane of the highway, right behind a large truck, surrounded by a car in the left lane and a motorcycle in the right lane. A large box suddenly falls from the truck's toward the subject, who then has to arbitrate between three alternatives: (1) keep straight and greatly endanger the car's passengers by hitting the box; (2) swerve to the left lane to avoid the box and hit the other car, moderately endangering all the passengers of the two vehicles; and (3) swerve to the right lane and hit the motorcyclist, severely endangering their life but with low harm to the passengers of the subject's car.

Philosophers and jurists agree that, whatever the driver's choice may be, neither their moral responsibility nor their legal liability is at stake here, their decision resulting from an instinctive reaction rather than a rational deliberate judgment. Now, by replacing the NAV by an AV, the results are an entirely different assessment of the decision taken by the algorithm in regard to both responsibility and liability. Unlike the driver who is only granted a few tenths of a second to understand the situation, make a decision, and implement it, the AV driving software will have a greater ability to react, as well as an a priori knowledge of the decision to take, its manufacturers having anticipated such scenarios and benefiting from an appropriate amount of time to identify the best alternative. Furthermore, we observe a shift in the decision maker's position; whereas the driver is directly involved in the dilemma situation, making a particular decision in praesenti to manage an existing scenario that may hurt their in the NAV case, manufacturers are indirectly involved in the dilemma situation, making general ex ante decisions to address potential scenarios that may endanger the passengers, but not them in the AV case.

To help conceptualize the problem, researchers have drawn an analogy between such AV dilemmas and the well-known "trolley problem" (e.g., Bonnefon et al., 2015; Lin, 2014; Loh & Loh, 2017; Sandberg & Bradshaw-Martin, 2013; Wallach & Allen, 2008), first conceived of by Foot (1967) and then notably explored by Thomson (1976, 1985). Inspired by the "trolley problem," Bonnefon et al. (2015) investigated the psychology of individuals faced with AV moral dilemmas. An initial survey completed on Amazon's Mechanical Turk platform found that although a large majority of participants were in favor of AVs programmed to minimize the number of total deaths, significantly fewer of them would actually buy such cars, understanding it could sacrifice them. The authors then concluded on the existence of a "social dilemma," summarized as: "People mostly agree on what should be done for the greater good of everyone, but it is in everybody's self-interest not to do it themselves" (p. 8). Assuming that a typical solution to overcome social dilemmas consists in regulators enforcing a targeted behavior, thus eliminating the opportunity to free ride, the researchers conducted another study focused on the impact of governmental regulation. This produced a paradoxical conclusion: while a regulation enforcing the so-called utilitarian AVs may solve the social dilemma, most people would likely disapprove it, ultimately leading to "a more serious problem," that is a conflict with the HMI: "regulation could substantially delay the adoption of AVs, which means that the lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether" (Bonnefon et al., 2016, pp. 1575–1576).

Supported by an extended team and for the purpose of collecting wide-scale information about individuals' preferences regarding AV moral dilemma, the researchers finally deployed the MM. Assuming an explicit foundation in Thomson's cases (based on the illustrations presented on its website), the online experimental platform offers trolley problem–type situations for which participants are asked to choose the less evil consequence between letting the car continue ahead or swerve into the other lane, resulting in different outcomes, implying at least one person's death. Each dilemma set contains 13 randomly selected situations designed to evaluate the participant's preferences according to nine factors.

The MM was greatly successful in its reach, collecting 39.61 million answers from 1.3 million respondents across 233 countries and territories in only 2 years. Its results reveal a global preference for sparing humans over animals, saving more lives versus fewer, and privileging the young versus the elderly. The researchers also identified three cultural clusters associated with different world regions and observed specific trends opposing individualistic and collectivistic cultures (Awad et al., 2018). Relayed by first-class international newspapers, the MM succeeded in reaching a much wider audience than the narrow sphere of AI ethicists, shedding some well-deserved light on to an important issue. However, the publishing of the MM results was soon followed by associated works pursuing antagonistic goals.

Of the Dangerous Uses of the MM Experiment

Elaborating on the results of the MM experiment, Awad declared that "What [they] are trying to show here is descriptive ethics: people's preferences in ethical decisions [. . .] But when it comes to normative ethics, which is how things should be done, that should be left to experts" (Vincent, 2018), firmly placing the MM in the spirit of Bonnefon et al.'s (2015) original works' ambition

> to be clear, we do not mean that these thorny ethical questions can be solved by polling the public and enforcing the majority opinion. Survey data will inform the construction and regulation of moral algorithms for AVs, not dictate them. (p. 4)

However, Awad and two other authors of the MM experiment cosigned another paper, published a month after the release of the MM results, presenting a "voting-based system for ethical decision making" (VBS) based on the MM data and arguing that "the starting point of [their] work was the realization that the MM dataset can be used not just to understand people, but also to automate decisions" (Noothigattu et al., 2018, p. 4). Building on the work of Greene et al. (2016) and Conitzer et al. (2017), the authors assert that "decision making can, in fact, be automated, even in the absence of such ground-truth principles, by aggregating people's opinions on ethical dilemmas." They present a "concrete approach for ethical decision making based on computational social choice" with the goal of "serving as a foundation for incorporating future ground-truth ethical and legal principles" which, once implemented on the MM data set, "can make credible decisions on ethical dilemmas in the autonomous vehicle domain" (Noothigattu et al., 2018, pp. 1, 2, 20). Let us examine two types of methodological limits regarding the MM data, justifying its disqualification to serve for any end other than descriptive ones, before demonstrating why computational social choice theories cannot be applied to ethical decisions.

Firstly, the scientific relevance of the VBS results is highly limited by the poor quality of the MM data, including strong selection biases across respondents: the "sample is self-selected" and "arguably close to the internet-connected, tech-savvy population that is interested in driverless car technology" (Awad et al., 2018, p. 63). Although Awad et al. try to minimize this bias's weight, based on the fact that the heterogeneity of answers across countries is correlated with cultural and economic specificities, the sample only takes into account the preferences of this "tech-savvy population," excluding all other people and in particular those reluctant to buy an AV, whose preferences, nonetheless, also deserve to be counted (they may end up in the "pedestrian" situation) and would certainly diverge. Furthermore, there are concrete reasons to be skeptical about the seriousness of many respondents when taking the MM tests, as well as the accuracy of their geolocalization, captured by their internet protocol address. No information is provided about eventual strategies to exclude virtual private networks users, while these are frequently used by a fourth of the internet population, a community expected to be overrepresented in the MM sample of tech-savvy people. At last, Awad et al. point out the simplistic aspect of the MM; it does not include uncertainty about consequences, thus implying risk management under limited information, whereas AVs' technologies are based on stochastic systems by nature probabilistic.

Secondly, whereas the MM project is explicitly presented as an applied trolley dilemma deriving from Thomson's cases, such an analogy encounters several objections. Two of them are presented by Nyholm and Smids (2016) regarding, on the one hand, the asymmetric number of interests at stake resulting from the normative ambition behind AV dilemmas and their expected multiple occurrences. On the other hand, the legal liability issues also tackled by AV dilemmas—whereas the trolley problem only questions moral responsibility, which significantly impacts decisions and constrains rights to action (Wood, 2011, pp. 74–75). Furthermore, these disanalogies do not preserve the MM from the criticisms expressed against the trolley problem itself, which essentially target the

inapplicability of its results. It has indeed been observed that respondents' decisions change with the level of concreteness of the experiment, being more reluctant to push the fat man over the bridge (Thomson's *fat man case*) in virtual reality (Francis et al., 2017), as well as to redirect an electroshock toward one mouse to avoid hitting five of them (Thomson's *bystander at the switch case*) in real conditions (Bostyn et al., 2018). Finally, it has also been remarked that the humoristic perception of the dilemma may alter respondents' decision-making process (C. W. Bauman et al., 2014).

Having established that the MM data cannot be used for normative intentions because of its methodological limits, I shall now expose two arguments demonstrating that the whole project to build an ethical decision-making system based on computational social choice theories, and upon which the VBS is based, is not only fallacious but also dangerous for our democracies.

First, let us recall than, by definition, only moral agents are capable of making moral decisions. Moral agents can be defined as autonomous subjects provided with a certain idea of the good and whose free will allows them to determine their own general principles of action from which to make particular decisions. They are capable of justifying them and responsible for the intended consequences. In contrast, there is today no algorithm autonomous in the philosophical sense and stochastic algorithms are particularly unable to justify each of their choices with consistent rules nor to be held accountable for the consequences of their actions. There is thus, for now, insufficient ground to question the moral status of such algorithms and also to refuse considering them as "moral proxies." Millar (2014) recalls that "a moral proxy is a person responsible for making healthcare decisions on behalf of another" when such a person is incapable to do so themselves (p. 128); the moral proxy of a moral agent is thus another *moral agent* making a *decision* in the best interests of the first one. This clarification is important because Conitzer et al. and Noothigattu et al. intend to create autonomous systems with the capacity to "make moral decisions" (Conitzer et al., 2017, p. 4831) or "make ethical decisions" (Noothigattu et al., 2018, p. 20), which jeopardize the traceability chain of decisions and responsibility, which in turn is necessary to fairly allocate sentences when algorithms produce harmful consequences. Consequently, because algorithms are not moral agents, they cannot make moral decisions and the production of ethical decisions cannot be automated.

Second, although the VBS does not produce moral judgments, it nevertheless claims to aggregate moral agents' judgments, which "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" (Conitzer et al., 2017, p. 4834). To refute this claim, let us focus on the approach underlying the MM. To solve an iconic problem of moral philosophy, Awad et al. opted for an unconventional approach, both psychological and descriptive. While the philosophical reasoning consists in transforming opinions into knowledge through a dialectical reflection and contradictory debate, the authors chose to infer general principles from aggregated a priori opinions, collected from individuals who have not received any background information to address these issues nor contradictors to challenge their answers. This choice to target people's *prima facie* perception of morality rather than reasoned and informed moral decisions results from the belief that, if philosophers have not been able to agree upon a solution yet, a consensus is not expected in the appropriate time (Rozieres, 2018). The importance of finding right decisions to fairly allocate harm in each dilemma situation is thus explicitly subordinated to the HMI (Shariff et al., 2017, p. 696). The fact that Awad et al. do not seek *moral rightness* or *fairness* across results, but the widest social acceptance, is also illustrated by the practical strategies they suggest to persuade people to buy AVs and solve the social dilemma, including virtue signaling and fear placebo (Shariff et al., 2017, p. 695).

However, people often change their minds about moral choices, whose volatility is highly and negatively correlated with the degree of information and deliberate reasoning they result from. Imagine a journalist asking people on the street about their perception of the ideal income tax rate, without informing them she was appointed by Congress to pilot a tax reform. Respondents may certainly give her a much lower rate than the present one. Not only are they abused by the journalist

who is hiding her survey's goal, but their answers do not even necessarily match their actual preferences. Once the survey is completed and the reform implemented, the same people might start complaining about the drastic loss of public services following the tax reform, arguing they would have changed their answers in favor of a higher tax rate had they been aware of the amount of public services these taxes were funding and taken more time to respond and had they been aware of the consequences of their replies. The MM faces the same issue because aggregating individual uninformed beliefs does not produce any common reasoned knowledge.

Noothigattu et al. (2018) actually concede that "Moral Machine users may be poorly informed about the dilemmas at hand, or may not spend enough time thinking through the options, potentially leading—in some cases—to inconsistent answers and poor models" but "believe, though, that much of this noise cancels out in Steps III [Summarization] and IV [Aggregation]" (p. 20), which is consistent with the idea that "idiosyncratic moral mistakes made by individual humans are washed out in the aggregate" (Conitzer et al., 2017, p. 4834). However, such aggregation does not reduce noise but only normalizes answers around an average social belief, one which presents no guarantee of approximating the right choices. If we define a wrong answer as a given answer which would change in the case that the respondent was given enough time, information, and opportunity to debate against a challenging opponent, then these "mistakes" are only washed out in the aggregate given two conditions: (a) the majority of people within the sample happen to be "right" (which is impossible to falsify) and (b) respondents are not consistent in their wrong answers (which they actually are). Whether people are right or wrong when prioritizing a category over another, they tend to stick to this rule across scenarios; they are "wrong" about the general principle upon which their answers are based but not about a particular answer.

The Instrumentalization of the Ethical Discourse

So far, I hope to have proven that using the MM data to develop a computational ethical decision-making system for normative ends such as the VBS is scientifically limited by the quality of the MM data. It is ontologically impossible because of the nature of such a system, which does not have the ability to make moral decisions. It is necessarily fallacious when aggregating uninformed beliefs to grant them an intersubjective common moral value. Finally, it is also dangerous, betraying the initial ambitions of the MM to close the public debate it allegedly intended to open, making an illegitimate use of data for the purpose of solving the AV dilemmas. However, there is one argument which could still be raised to justify the use of VBS-like systems, suggesting inadequate but not too shocking solutions to AV dilemmas in order to accelerate their deployment—and that is the HMI.

Noothigattu et al. (2018) assert that

> in their work on fairness in machine learning, Dwork et al. concede that, when ground-truth ethical principles are not available, we must use an 'approximation as agreed upon by society.' But how can society agree on the ground truth—or an approximation thereof—when even ethicists cannot? (p. 1)

This justification is to be refuted on three levels: firstly, because the work quoted has very little relevance for ethical considerations; secondly, because there is in fact a ground upon which ethicists do agree; and thirdly, because the underlying axiom justifying the need to develop an ethical decision-making system in the absence of univocal agreement about any ground truth is unacceptable.

At first, the work cited by Noothigattu et al. is supported by a poor theoretical grounding, merely mentioning a short definition of "equality of opportunity" proposed by Rawls, given out of context and without any further comment regarding Rawls's theory of justice (Dwork et al., 2012, p. 3). In

addition, the paper written by researchers at Microsoft Research is clearly not ethics oriented but business oriented (p. 1):

> In keeping with the motivation of fairness in online advertising, our approach [. . .] allows the vendor to benefit from investment in data mining and market research in designing its classifier, while our absolute guarantee of fairness frees the vendor from regulatory concerns.

Secondly, it is true that philosophers still debate the priority between the moral obligation not to infringe individuals' rights and the moral permission to seek to save more lives rather than fewer lives. They however tend to agree on the other aspects of the dilemma, especially refusing unfairly discriminatory criteria, and mostly differ in their interpretation of the theoretical problem rather than on the principles upon which the decision should be made. Furthermore, there exist several ways to settle such disagreements, among them the law production process that enables people to "agree to disagree" in modern democracies. When investigating the dilemma to build a common identity in a multicultural state, Taylor (2017) recognizes the challenge to combine the need for strong popular cohesion around a common political identity to develop social trust, with a multiculturalist condition to avoid the exclusion of minorities. Taylor comes to the conclusion that democratic regimes should be such that citizens are free because they not only take part in the decision-making unit by a vote equal to that cast by others, but because they are also included within a fair common discussion preceding the vote. Whereas the VBS is based on the assumption that training an algorithm on a sample of collected a priori uninformed inclinations to identify the compromise with the greatest chances to be accepted by the population may be a quicker way to bring AVs on roads—rather than waiting for the outcome of a public debate—it could actually lead to the opposite result. In fact, while most people could be in favor of a principle that seeks to save more lives rather than fewer lives, they may nonetheless reject it for the sole reason that it results from a procedure perceived imposed rather than legitimate, just like a court is often obliged to reject useful but unacceptable evidence when its sourcing is irregular.

Thirdly, it could be argued that some ends justify all means and that an early deployment of AVs would be one of them. This is the argument of HMI's backers, who include officials and manufacturers, such as Mark Rosekin and Tesla's CEO Elon Musk, but also academics such as Bonnefon et al. (2015, p. 2). The HMI could be enunciated as follows: the fact that thousands of people die every year on the roads due to poor human driving skills justifies the existence of a moral obligation for car manufacturers to deploy AVs as soon as possible, and for regulators to authorize their marketing as early as AVs can reduce the net balance of total annual death by one (Kalra & Groves, 2017), to implement regulation with low liability for manufacturers in case of accident in order to avoid discouraging them (Bonnefon et al., 2015; Hevelke & Nida-Rümelin, 2015), and even to prohibit the use of NAVs when AVs become safer than them (Sparrow & Howard, 2017).

Let us firstly agree on the fact that HMI's backers would need to explain either why car accidents are a less tolerable cause of death than starvation or why hungry people's lives in developing countries are worth less than healthy people's in a developed country (even if not born yet). Much more significant and assured results in terms of life saving could indeed be reached by investing AVs' research and development budgets to feed the 821 million undernourished people worldwide (Food and Agriculture Organization, 2018, p. 2). Another argument against the claim that delaying AVs results in sacrificing lives is given by Lin (2013) who relates the issue to Parfit's (1986) nonidentity problem. Although AVs may halve the number of deaths due to car accidents, Lin says the people who will still die will unlikely all be the same ones as those who would have died otherwise. In other words, the net total of lives "saved" would remain positive if the introduction of AVs provoked 999 additional deaths while preventing 1,000, resulting in changing the identity of many victims. The equivalence of deaths presupposed by utilitarianists may then be challenged

when considering the net average level of responsibility. Assuming that 94% of NAVs accidents are caused by human error (i.e., almost all NAV accidents involve at least one person with some degree of responsibility), whereas 100% of AV accidents involve at least some fully innocent people (the AV passengers), we may rationally suppose that a major part of the 999 traded dead people would be less responsible for their own death than a majority of the 1,000. In fact, we surely concede that it would be unfair to save an at fault drunk-driving person's life at the cost of AV passengers' lives.

Hence, my point here is not that AVs deployment should be unnecessarily delayed, but that the instrumental use of moral considerations as leverage to develop a favorable regulation for manufacturers has no solid foundations. In contrast, the duty for governments to only allow AVs if they are implemented with fair moral principles to answer dilemma situations cannot be subordinated to the opportunity offered by AVs to save a number of people's lives when engaged in a driving activity they know to entail perils. As summed up by the German Ethics Commission, "there is no ethical rule that always places safety before freedom" (Federal Ministry of Transport and Digital Infrastructure, 2017, p. 20), and it would be wrong to believe that an old woman's rights would only be infringed if she happens to be involved in a dilemma situation where the AV is instructed to drive over her instead of a young boy because she is elder. They would be violated every single day from the legal deployment of AVs implemented with such preferences, and she would be aware that her life is valued as less worthy than any younger person in society.

Finally, it is tempting to fall in the trap of considering the MM without the VBS as a valuable experiment. Let us consider two strong negative effects it had on public opinion that prove it wrong.

First, the principle of the MM suggests that individuals' value of life varies with their characteristics and that the nine differentiation factors selected by Awad et al. (including sparing men vs. women, the fit vs. the less fit, and those with higher social status vs. others) are relevant to conduct life arbitrations. As demonstrated somewhere else (Etienne, in press), not only are most of these criteria morally irrelevant, but the whole MM's characteristic-based approach is dangerous in itself, polarizing the debate around erroneous AV dilemmas. The concrete damage deriving from the MM is then psychological, enforcing people's belief that it is acceptable and morally relevant to allocate death based on gender, weight, or social status.

Second, a side effect of the MM popularity was to distract from other first-order ethical issues such as preserving the integrity of embedded systems against hacking and threats of AV use for terrorist ends, losing the possibility to react in critical situations (e.g., exceeding the authorized speed limit for medical emergencies or to escape an impending aggression), or addressing the impact of AVs on the job market. Two of them which have received particularly low attention are the probable forthcoming prohibition of NAVs—announced by Bill Gates and Elon Musk (Dredge, 2015) and which is necessary to release the full extent of AVs' expected benefits, notably improving traffic fluidity with intersection traffic management algorithmic regulators (Au & Stone, 2010)—and the building of an AV-based mass surveillance system. AVs are equipped with an exhaustive range of sensors including internal and external cameras, capturing a flow of information for which public authorities have already declared their interest (European Commission, 2018, p. 13), thus calling for strong data protection.

Conclusion

AVs are equipped with several of the most promising applications in AI, and their development will result in profound ethical, social, political, and economic impacts on the lives of billions of people. They can deservedly be considered as the ethical challenge of the decade in AI ethics, in the sense that the way their underlying issues will be approached and settled will certainly mark jurisprudence, giving a direction to the development of the discipline. This is precisely why it is important to resist the sirens of the market calling for emergency responses. Computational approaches should be

deployed with greater prudence to inform human choices rather than to substitute them. Here again, there is a high risk of ceding to the temptation of using them to solve complex social decisions, short-circuiting public consultation and producing an irresponsible nonhuman ethics, incapable of consistently explaining its choices, and justifying its legitimacy. Such a threat is ironically captured by the fresco of Cesare Maccari chosen by Noothigattu et al. to illustrate the VBS project's webpage (https://www.media.mit.edu/projects/a-voting-based-system-for-ethical-decision-making/overview/), which at first glance depicts an orator discoursing in front of a chamber of representatives but actually represents Cicero denouncing the Catiline's plotting to the Senate and its dangers for the Roman republic.[2]

Data and Software Information

Link to the Moral Machine project (https://www.media.mit.edu/publications/the-moral-machine-experiment/).
Link to the Moral Machine platform (http://moralmachine.mit.edu).

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. By autonomous vehicle, let us refer to a vehicle with either Level 4 or Level 5 automation, according to the Society of Automotive Engineers' classification (https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety), when the self-driving mode is activated.
2. Cicerone denuncia Catilina, fresco of Cesare Maccari, 1889, Roma, Palazzo Madama, Roma.

References

Au, T., & Stone, P. (2010). Motion planning algorithms for autonomous intersection management. In *Proceedings of the 1st AAAI Conference on Bridging the Gaps Between Task and Motion Planning* (pp. 2–9). AAAI Press.

Awad, E., Sohan, D., Richard, K., Jonathan, S., Joseph, H., Azim, S., Jean-François, B., & Iyad, R. (2018). The moral machine experiment. *Nature*, *563*, 59–64.

Bauman, C. W., Peter, M. A., Daniel, B. M., & Caleb, W. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*, 536–554.

Bauman, M., & Youngblood, A. (2017). *Why waiting for perfect autonomous vehicles may cost lives*. RAND Corporation.

Bonnefon, J., Azim, S., & Iyad, R. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? ArXiv. https://128.84.21.199/abs/1510.03346v1

Bonnefon, J., Azim, S., & Iyad, R. (2016). The social dilemma of autonomous vehicles. *Science*, *352*, 1573–1576.

Bostyn, D. H., Sybren, S., & Arne, R. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, *29*, 1084–1093.

Conitzer, V., Walter, S., Jana, B. S., Yuan, D., & Max, K. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 4831–4835). AAAI Press.

Dredge, S. (2015). Elon Musk: Self-driving cars could lead to ban on human drivers. *The Guardian*. https://www.theguardian.com/technology/2015/mar/18/elon-musk-self-driving-cars-ban-human-drivers

Dwork, C., Moritz, H., Toniann, P., Reingold, T., & Rich, Z. S. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). Association for Computing Machinery.

Etienne, H. (in press). A practical role-based approach to solve moral dilemmas for self-driving cars.

European Commission. (2018). *On the road to automated mobility: An EU strategy for mobility of the future* (p. 283). COM.

Federal Ministry of Transport and Digital Infrastructure. (2017). *Ethics commission on automated and connected driving* [Report]. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Food and Agriculture Organization. (2018). *The state of food security and nutrition in the world*. United Nations.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Francis, K. B., Charles, H., Ian, H. S., Michaela, G., Giorgio, G., Grace, A., & Sylvia, T. (2017). Virtual morality: Transitioning from moral judgment to moral action? *PLoS One*, *12*, e0171793.

Greene, J., Francesca, R., John, T., Kristen, V. B., & Brian, W. (2016). Embedding ethical principles in collective decision support systems. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (pp. 4147–4151). Association for Computing Machinery.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics*, *21*, 619–630.

Kalra, N., & Groves, D. G. (2017). *The enemy of good: Estimating the cost of waiting for nearly perfect automated vehicles*. RAND Corporation.

Lin, P. (2013). The ethics of saving lives with autonomous cars is far murkier than you think. *Wired*. https://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/

Lin, P. (2014). Here's a terrible idea: Robot cars with adjustable ethics settings. *Wired*. https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/

Lin, P. (2015). The ethical dilemma of self-driving cars. *Ted-Ed*. https://www.ted.com/talks/patrick_lin_the_ethical_dilemma_of_self_driving_cars?language=en

Loh, W., & Loh, J. (2017). Autonomy and responsibility in hybrid systems: The example of autonomous cars. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0* (pp. 35–50). Oxford University Press.

Millar, J. (2014). Technology as moral proxy: Autonomy and paternalism by design. In *Proceedings of the 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*. Association for Computing Machinery.

Mohr, D., Hans-Werner, K., Paul, G., Dominik, W., & Timo, M. (2016). *Automotive revolution-perspective towards 2030. How the convergence of disruptive technology-driven trends could transform auto industry*. McKinsey & Company.

Noothigattu, R., Snehalkumar, G., Neil, S., Edmond, A., Sohan, D., Iyad, R., Pradeep, R., & Ariel, P. D. (2018). A voting-based system for ethical decision making. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*. AAAI Press.

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

Parfit, D. (1986). *Reasons and persons*. Oxford University Press.

Rozieres, G. (2018). Les voitures autonomes doivent-elles vous sacrifier pour sauver un enfant ou un chien? [Should autonomous cars sacrifice you to save a child or a dog?] *The Huffington Post*. https://www.huffingtonpost.fr/2018/10/24/les-voitures-autonomes-doivent-elles-vous-sacrifier-pour-sauver-un-enfant-ou-un-chien_a_23570383/

Sandberg, A., & Bradshaw-Martin, H. (2013). What do cars think of trolley problems: Ethics for autonomous cars? In *Proceedings of the 2013 International Conference, Beyond AI*. Springer.

Shariff, A., Jean-François, B., & Iyad, R. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*, 694–696.

Singh, S. (2015). *Critical reasons for crashes investigated in the National motor vehicle crash causation survey [Traffic Safety Facts Crash Stats. Report No. DOT HS 812 115*]. National Highway Traffic Safety Administration.

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies*, *80*, 206–215.

Taylor, C. (2017, October 14). Political identity and the problem of democratic exclusion. *ABC Religion and Ethics*. http://www.abc.net.au/religion/articles/2016/04/29/4452814.htm

Thomson, J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*, 204–217.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Vincent, J. (2018). Global preferences for who to save in self-driving car crashes revealed. *The Verge*. https://www.theverge.com/2018/10/24/18013392/self-driving-car-ethics-dilemma-mit-study-moral-machine-results

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wood, A. (2011). Humanity as an end in itself. In D. Parfit & S. Scheffler (Eds.), *On what matters* (Vol. *2*, pp. 58–82). Oxford University Press.

World Health Organization. (2018). *Global status report on road safety 2018*. https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

Author Biography

Hubert Etienne is completing his doctoral degree between Ecole Normale Supérieure, Philosophy department, Sorbonne University, Sciences & Engineering department, and Facebook AI Research. He is research associate at the Centre for Technology & Global Affairs at Oxford University and lecturer in AI Ethics at ESCP Europe.