Law, Innovation and Technology

ISSN: 1757-9961 (Print) 1757-997X (Online) Journal homepage: www.tandfonline.com/journals/rlit20

The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles

Hubert Etienne

To cite this article: Hubert Etienne (2021) The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles, Law, Innovation and Technology, 13:1, 85-107, DOI: 10.1080/17579961.2021.1898310

To link to this article: https://doi.org/10.1080/17579961.2021.1898310

LAW, INNOVATION AND TECHNOLOGY
2021, VOL. 13, NO. 1, 85–107
https://doi.org/10.1080/17579961.2021.1898310

Routledge
Taylor & Francis Group

The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles

Hubert Etienne

Department of Philosophy, Ecole Normale Supérieure, Paris, France; Laboratory of Computer Sciences, Sorbonne University, Paris, France

ABSTRACT

This paper reveals the dangers of the Moral Machine 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 and provides reasons why it is inadequate in supporting ethical and juridical discussions to determine the moral settings for autonomous vehicles. Demonstrating the inner fallacy behind computational social choice methods when applied to ethical decision-making, it also warns against the dangers of computational moral systems, such as the 'voting-based system' recently developed out of the Moral Machine's data. Finally, it discusses the Moral Machine's ambiguous impact on public opinion; on the one hand, laudable for having successfully raised global awareness with regard to ethical concerns about autonomous vehicles, and on the other hand pernicious, as it has led to a significant narrowing of the spectrum of autonomous vehicle ethics, de facto imposing a strong unidirectional approach, while brushing aside other major moral issues.

ARTICLE HISTORY Received 23 July 2019; Accepted 8 May 2020

KEYWORDS Autonomous vehicles; self-driving cars; AI ethics; Moral Machine; trolley problem

1. Introduction

Whether they reach level 4 or level 5 autonomy within the next decade, autonomous vehicles (AV)[1] – whose tremendous expected benefits justify the fierce competition between both manufacturers and national governments – represent a crucial challenge for AI ethics. While players have

CONTACT Hubert Etienne hubert.etienne@sciencespo.fr Department of Philosophy, Ecole Normale Supérieure, 45 rue d'Ulm, 75005, Paris, France; Laboratory of Computer Sciences, Sorbonne University, 4 place Jussieu, 75005, Paris, France

[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 (www.nhtsa.gov/technology-innovation/automated-vehicles-safety), when the self-driving mode is activated. Unless indicated otherwise, all websites were (accessed 1 December 2019).

This article has been republished with minor changes. These changes do not impact the academic content of the article.

invested up to $80 billion between August 2014 and June 2017[2] in the hope of conquering a solid share of a market forecasted to reach $6.7 trillion by 2030,[3] governments have understood the necessity of adapting national regulations to support self-driving testing, foreseeing AVs' high potential to contribute to economic growth and increased public safety.[4] Such advances are, however, not without their own legal and ethical issues. The most discussed so far has been that of moral responsibility and legal liability in the case of fatal accidents, which became a tragic reality in March 2018, when a self-driving Uber car killed a pedestrian in Arizona.[5]

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'[6] a theoretical framework to address the resulting moral dilemmas. Awad et al.[7] 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 great success of the experiment, gathering about 40 million answers, then became the starting point for Noothigattu et al.[8] to develop a 'voting-based system' (VBS), grounded in computational social choice theories, with the intention of automating ethical decisions by aggregating the individuals' moral preferences collected by the MM.

This paper presents a multi-level 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 the use of the MM data for normative purposes, and the inner fallacy of attempting to automate ethical decision-making processes. The first part analyses the construction of the AV moral dilemmas and its current approach in the debate, from the advent of a moral imperative supporting the development of AVs to the conceptualisation of the dilemmas on the trolley problem's model and the deployment of the MM. The second part then criticises the use of the MM data by the VBS and refutes the possibility of developing a

[2]Cameron F Kerry and Jack Karsten, 'Gauging Investments in Self-Driving Cars' *Brookings* (16 October 2017), www.brookings.edu/research/gauging-investment-in-self-driving-cars/ (accessed 30 December 2020).

[3]Detlev Mohr and others, *Automotive Revolution-Perspective Towards 2030. How the Convergence of Disruptive Technology-Driven Trends Could Transform Auto Industry* (McKinsey & Company, 2016) 6.

[4]European Commission (EC), *On the Road to Automated Mobility: An EU Strategy for Mobility of the Future* COM(2018) 283, 2.

[5]Sam Levin and Julia C Wong, 'Self-Driving Uber Kills Arizona Woman in First Crash Involving Pedestrian', *The Guardian* (London, 19 March 2018).

[6]Philippa Foot, 'The Problem of Abortion and the Doctrine of Double Effect' (1967) 5 *Oxford Review* 5; Judith J Thomson, 'Killing, Letting Die, and the Trolley Problem' (1976) 59(2) *The Monist* 204; Judith J Thomson, 'The Trolley Problem' (1985) 94(6) *Yale Law Journal* 1395.

[7]Edmond Awad and others, 'The Moral Machine Experiment' (2018) 563 *Nature* 59.

[8]Ritesh Noothigattu and others, 'A Voting-based System for Ethical Decision Making' (2018) *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

legitimate coherent computational system to automate ethical decisions. Finally, the last part denounces an instrumentalization of the ethical discourse, rejecting the 'highest moral imperative' underlying this project, and exposing the distracting effects of the MM on public opinion, having both polarised the debate on erroneous principles and categories, and prevented relevant ethical issues from receiving appropriate attention.

2. A responsibility issue resulting from the transfer of autonomy

This first section analyses each step of the construction of the AV moral dilemmas, from the advent of a moral imperative supporting the development of AVs to the conceptualisation of such a dilemma on the trolley problem's model and the deployment of the MM. It insists on the economic and political issues, which I hold to play a major role in the justification of this moral imperative referred to as the HMI and further discussed in Part 4.

2.1. From economic and social benefits toward a moral imperative to develop autonomous vehicles

According to their proponents, AVs represent a great advancement in circulating people and goods, expected to provide the automotive industry with tremendous business opportunities, customers with substantive advantages, and benefits to society as a whole in regards to road safety, economic growth and traffic management.

For the automobile industry, AVs are the centre of a common vision to expand the market and deeply modify its structure, specifically regarding its revenue streams. Taking automotive competition beyond its traditional differentiation factors, the AV project aims to produce a radical change in mobility consumption behaviour, from a private car ownership model to a system dominated by shared mobility solutions.[9] Partnerships between technology companies and car manufacturers such as Waymo and Fiat Chrysler, Uber and Volvo, or Lyft and General Motors, are thus leading the race to complete the disruption of the mobility sector initiated by Uber a decade ago. These companies not only expect overwhelming revenue increases, as the aggregate of both price and volume effects induced by the replacement of human drivers but also a vast diversification in their services to absorb the whole of the delivery economy.

[9]Mohr (n 3) 8; Wolfgang Gruel and Joseph M Stanford, 'Assessing the Long-Term Effects of Autonomous Vehicles: A Speculative Approach' (2016) 13 *Transportation Research Procedia* 28. While the McKinsey study demonstrates that a change in mobility behaviour will make economic sense, Gruel and Stanford explain why such change may be required to attain a greener and more sustainable mobility system, which the introduction of AVs alone cannot achieve.

On the consumers' side, one can expect safer rides under all circumstances, allowing people to go home safely whatever their alcohol or tiredness level, as well as better time management, enabling them to rest, work or conduct any other activity during their journey. AVs also promise to be more economical for both shared-ride users and car owners. The average operating cost of an electric-powered vehicle was estimated as 2.3x lower than a fuel-powered vehicle,[10] thus compensating for the sale price premium in a few years, which is itself expected to decrease with AVs democratisation. Furthermore, AVs claim more inclusive mobility, opening the roads to people with reduced driving capacity (e.g. those with mobility or visual impairments), together with all those who do not have a driver's license.

With reference to governments and societies, the deployment of AVs is promoted as leading to better traffic management, a significant positive environmental impact, and considerable public savings. In mixed-autonomy traffic, a low proportion of AVs may be sufficient to increase traffic fluidity by reducing congestion and increasing the average speed – some suggest that traffic comprised of only 10% AVs could even double the average car speed[11] – while cutting carbon dioxide emissions by up to 60%[12] due to greater fuel efficiency and speed management. In fully autonomous traffic, most road infrastructure (signage, radars, etc.), together with their dedicated agents (traffic police, public transportation drivers, etc.), would become redundant, resulting in massive cost reductions for public administrations. A recent study also suggests that the introduction of AVs in Boston, USA, could reduce the need for parking spaces up to 50% by 2030.[13]

Finally, the greatest advantage of AVs decidedly dwells in its potential to make roads safer. Around 1.35 million people die every year in traffic accidents across the world[14] – including 25,500 people in the EU[15] and 37,461 in the US[16] in 2016 – while 94% of car accidents are said to result from human error[17] AVs are expected to drastically reduce this number, matching

[10]Michael Sivak and Brandon Schoettle, 'Relative Costs of Driving Electric and Gasoline Vehicles in the Individual U.S. States' (2018) Report No. SWT-2018-1, University of Michigan, 6.
[11]Eugene Vinitsky and others, 'Benchmarks for Reinforcement Learning in Mixed-Autonomy Traffic' (2018) 87 *Proceedings of the 2nd Conference on Robot Learning*, PMLR 11.
[12]Michele Bertoncello and Dominik Wee, *Ten Ways Autonomous Driving Could Redefine the Automotive World* (McKinsey & Company, 1 June 2015).
[13]Nikolaus S Lang and others, *Making Autonomous Vehicles a Reality. Lessons from Boston and beyond* (The Boston Consulting Group, October 2017).
[14]World Health Organization, *Global Status Report on Road Safety 2018*, 4.
[15]European Commission, *Statistiques de la sécurité routière pour 2016: que révèlent les chiffres?* (10 April 2018) Information (n 1).
[16]National Center for Statistics and Analysis, *2016 Fatal Motor Vehicle Crashes: Overview* (2017) Traffic Safety Facts Research Note. Report No. DOT HS 812 456, National Highway Traffic Safety Administration, 1.
[17]Santokh Singh, *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey* (2015) Traffic Safety Facts Crash Stats. Report No. DOT HS 812 115, National Highway Traffic Safety Administration, 1.

a top political priority of both the US National Safety Council, which leads the 'Road to Zero Coalition', and the EU, whose 'Zero Vision' aims for zero fatalities on European roads by 2050.[18] It is worth underlining the fact that road mortality constitutes a considerable economic burden on nation states' economic growth and healthcare systems. The National Highway Traffic Safety Administration (NHTSA) estimates the impact of vehicle crashes amounts to $242 billion in US economic activity, with an additional $594 billion due to injuries.[19] 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,[20] supporting relaxed regulations for manufacturers and lowering their liability in case of accident to avoid discouraging them.[21] Let us refer to this claim as the 'highest moral imperative' (HMI), particularly well-illustrated by Mark Rosekin's remarks while still chief regulator at the NHTSA: '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?'.[22] 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, which leads them to develop simulation instruments designed to help policy makers identify this critical moment,[23] while others assert, on the same ground, that regular cars should be prohibited as soon as AVs become safer.[24]

2.2. The birth of a responsibility issue and the Moral Machine experiment

Although AVs supporters advance a reasonable argument in favour of early deployment, some advocate that additional issues should first be fully 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 realistically prevent all accidents, which could still occur through

[18]European Commission, *Roadmap to a Single European Transport Area – Towards a Competitive and Resource Efficient Transport System* COM(2011) 144, 22.
[19]National Highway Traffic Safety Administration, 'Automated Vehicles for Safety', https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety (accessed 30 December 2020).
[20]Jean-François Bonnefon, Azim Shariff and Iyad Rahwan, 'The Social Dilemma of Autonomous Vehicles' (2016) 352(6293) *Science* 1575; Azim Shariff, Jean-François Bonnefon and Iyad Rahwan, 'Psychological Roadblocks to the Adoption of Self-Driving Vehicles' (2017) *Nature Human Behaviour* 694.
[21]Alexander Hevelke and Julian Nida-Rümelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21(3) *Science and Engineering Ethics* 619, 629.
[22]Melissa Bauman and Alyson Youngblood, 'Why Waiting for Perfect Autonomous Vehicles May Cost Lives' (RAND Corporation, 2017).
[23]Nidhi Kalra and David G Groves, *The Enemy of Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles* (RAND Corporation, 2017).
[24]Robert Sparrow and Mark Howard, 'When Human Beings are Like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport' (2017) 80 *Transportation Research Part C: Emerging Technologies* 206, 209–10.

technical outages or meteorological conditions, sometimes resulting in complex situations where no evident choice could be unanimously preferred. One of these is analysed by Patrick Lin[25] in a thought experiment where an agent is driving a non-autonomous 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 towards the agent who does not have sufficient space to stop the car safely, and thus needs 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; (3) swerve to the right lane and hit the motorcyclist, severely endangering his or her life but with low harm to the passengers of the subject's car.

This illustrates the sort of complex situation a driver (let us call her Aliénor) may encounter, facing a moral dilemma involving several people's lives, including her own, with no evident unequivocal solution. Under such conditions, moral philosophers tend to agree that, whatever Aliénor's choice may be, neither her moral responsibility nor her legal liability is at stake here – unless she entered this situation by breaking the law, for instance by exceeding the authorised speed limit – as her decision results from an instinctive reaction, rather than a rational deliberate judgment. Now, by replacing the NAV with an AV, the results are an entirely different assessment of the decision taken by the algorithm in regards to both responsibility and liability. Unlike Aliénor who is only granted a few tenths of a second to both understand the situation, make a decision and implement it, the AV driving software has a much better ability to react, as well as an *a priori* knowledge of the appropriate 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 Aliénor is directly involved in the dilemma situation, making a particular decision *in praesenti* to manage an existing scenario that may hurt her in the NAV case, manufacturers are indirectly involved in the dilemma situation, making general *ex ante* decisions to address potential scenarios that may endanger Aliénor, but not them in the AV case.

To help conceptualise the problem, researchers have drawn an analogy between such AV dilemmas and the well-known 'trolley problem',[26] first

[25]Patrick Lin, 'The Ethical Dilemma of Self-Driving Cars' (2015) Ted-Ed.

[26]Examples include Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2008); Anders Sandberg and Heather Bradshaw-Martin, 'What Do Cars Think of Trolley Problems: Ethics for Autonomous Cars?', *Proceedings of the 2013 International Conference*, Beyond AI, 2013; Noah J Goodall, 'Machine Ethics and Automated Vehicles' in G Meyer and S Beiker (eds), *Road Vehicle Automation* (Springer, 2014) 93. Patrick Lin, 'Here's a Terrible Idea: Robot Cars With Adjustable Ethics Settings' *Wired* (18 August 2014); Jean-François Bonnefon, Azim Shariff and

conceived of by Philippa Foot[27] and then notably explored by Judith Thomson[28] and Frances Kamm.[29] Foot's thought experiment questions which is the lesser of two evil actions to make for a moral agent when confronted with a critical situation, traditionally addressed either from a deontological view (considering that killing is worse than letting die), or from a consequentialist approach (for which saving five people is better than saving one).

Inspired by the trolley problem, Bonnefon et al. investigated the psychology of individuals faced with AV moral dilemmas. An initial study, comprised of three surveys completed on Amazon's Mechanical Turk platform, found that although a large majority of participants (c. 75%) were in favour of 'utilitarian AVs' (i.e. cars programmed to minimise the number of total deaths), significantly fewer believed that AVs would actually be programmed to this end, foreseeing the incentive for manufacturers to prioritise the life of AV passengers over others' lives. Furthermore, despite feeling comfortable with others buying utilitarian AVs, respondents were much less willing to buy such cars themselves. The authors then concluded on the existence of a 'social dilemma', summarised 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'.[30] Assuming that a typical solution to overcome social dilemmas consists of regulators enforcing a targeted behaviour, thus eliminating the opportunity to free ride, the researchers conducted another study, focused on the impact of governmental regulation. From the analysis of six surveys also submitted to the Mechanical Turk platform, they came out with a paradoxical conclusion: 'regulation may be necessary, but at the same time counterproductive'.[31] In fact, while regulation may solve the social dilemma, they find that most people would likely disapprove of a regulation that would enforce utilitarian AVs, 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'.[32]

Iyad Rahwan, 'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?' (2015) *ArXiv*; Christian Gerdes and Sarah Thornton, 'Implementable Ethics for Autonomous Vehicles', in M Maurer, C Gerdes, B Lenz and H Winner (eds), *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte* (Springer, 2015) 87; Giuseppe Contissa, Francesca Lagioia and Giovanni Sartor, 'The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law' (2017) 25(3) *Artificial Intelligence and Law* 365; Wulf Loh and Janina Loh, 'Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars', in P Lin, K Abney and R Jenkins (eds), *Robot Ethics 2.0* (Oxford University Press, 2017) 35.

[27]Foot (n 6).

[28]Thomson (n 6).

[29]Frances M Kamm, 'Harming Some to Save Others' (1989) 57(3) *Philosophical Studies* 227.

[30]Bonnefon et al. (n 26) 8.

[31]Bonnefon et al. (n 20) 1575.

[32]Ibid. 1575–6.

Supported by other colleagues for the purpose of collecting wide-scale information about individuals' preferences regarding AV moral dilemmas, identifying potential cultural regularities and mapping geographical trends, the researchers finally deployed the MM. Assuming an explicit foundation in Thomson's cases – based on the illustrations presented on their 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 thirteen randomly selected situations designed to evaluate the participant's preferences according to nine factors: (1) sparing humans versus pets, (2) staying on course versus swerving, (3) sparing passengers versus pedestrians, (4) sparing more lives versus fewer lives, (5) sparing men versus women, (6) sparing the young versus the elderly, (7) sparing pedestrians who cross legally versus jaywalking, (8) sparing the fit versus the less fit, and (9) sparing those with higher social status versus those with lower social status.

The MM was undeniably successful in its reach, collecting 39.61 million answers from 1.3 million respondents across 233 countries and territories in only two 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 (Western, Eastern, Southern), and observed specific trends opposing individualistic and collectivistic cultures.[33] 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 an issue that would otherwise have remained in the shadow of the innovation race.[34] However, the publishing of the MM results was soon followed by associated works pursuing antagonistic goals.

3. Of the dangerous uses of the Moral Machine experiment

In this Part of the article, I discuss the problematic shift in approach from the MM to the VBS, criticising the disloyal intentions it reveals about the whole experiment, presenting two main types of methodological limitations disqualifying the use of MM data for such normative intentions, as well as several arguments against the relevance of computational social choice theories to provide satisfying answers to ethical dilemmas.

[33] Awad et al. (n 7).

[34] The words of Christoph von Hugo, Mercedes-Benz's manager of driver assistance systems and active safety, at the 2016 Paris Motor Show proves that many manufacturers would have otherwise been keen to solve AV moral dilemmas on their own: 'If you know you can save at least one person, at least save that one. Save the one in the car' (Lindsay Dodgson, 'Why Mercedes Plans to Let Its Self-Driving Cars Kill Pedestrians in Dicey Situations' *Business Insider France* (Paris, 12 October 2016)).

3.1. From the Moral Machine to the voting-based system

Elaborating on the results of the MM experiment, Edmond 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',[35] firmly placing the MM in the spirit of Bonnefon et al.'s 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'.[36] However, Awad and two other authors of the MM experiment co-signed another paper, published a month after the release of the MM results, presenting a 'voting-based system for ethical decision making' (VBS) trained on the MM dataset 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'.[37] In the wake of Joshua Greene's connection between computational social choice and ethical decision making[38] together with Vincent Conitzer's statement that aggregating moral views could lead to the development of a morally better system,[39] 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 dataset, 'can make credible decisions on ethical dilemmas in the autonomous vehicle domain'.[40] The shift in the researchers' intentions regarding the collected data between the MM and the VBS cannot seriously be associated to anything else than a clear ethical fault and an academic fellony against the experiment's subjects. It should be added to the list of scandals faced by the MIT Media[41] Lab and leading us to challenge the sincerity of the authors' philanthropic ambitions.

[35] James Vincent, 'Global Preferences for Who to Save in Self-Driving Car Crashes Revealed' *The Verge* (24 October 2018).
[36] Bonnefon et al. (n 26) 4.
[37] Noothigattu et al. (n 8) 4.
[38] Joshua Greene and others, 'Embedding Ethical Principles in Collective Decision Support Systems' (2016) *Proceedings of the 30th AAAI Conference on Artificial Intelligence* 4147–51.
[39] Vincent Conitzer and others 'Moral Decision Making Frameworks for Artificial Intelligence' (2017) *Proceedings of the 31st AAAI Conference on Artificial Intelligence* 4831–5.
[40] Noothigattu et al. (n 8) 1, 2, 2, 20.
[41] It may be relevant to point out here that between the submission and the final acceptance of the present article, the MIT Media Lab (ML) – from where the MM and the VBS experiments were conducted – went through a series of scandals. These involve undisclosed funding received from Jeffrey Epstein (Noam Cohen 'Dirty Money and Bad Science at MIT's Media Lab' (*Wired*, 16 January 2020)) to the benefit of both the ML and its director's investment funds, resulting in the latter's resignation and public apology (www.media.mit.edu/posts/my-apology-regarding-jeffrey-epstein/), as

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' justifying the caution that 'policymakers should not embrace [this] data as the final word on societal preferences'.[42] Although Awad et al. then try to minimise this bias's weight, based on the fact that the heterogeneity of answers across countries exhibits cultural and economic specificities, as a matter of fact, their sample only takes into account the preferences of this 'tech-savvy population'. The MM data is then certainly biased in favour of a population more likely to buy AVs, thus to occupy the passenger seat in the dilemma, rather than those more susceptible to end up in the pedestrian position. These latter's preferences, nonetheless, also deserve to be counted and may 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 geo-localization, captured by their IP address. In fact, no information is provided about eventual strategies to exclude virtual private networks users – VPNs being frequently used by 26% of the internet population[43] – a community expected to be overrepresented in the MM sample of tech-savvy people. While Awad et al. point out the simplistic aspect of the MM, this does not include uncertainty about consequences, thus implying risk management under limited information, a flaw that also makes the MM technologically unviable.[44]

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 Sven Nyholm and Jilles Smids.[45] Firstly, there is an asymmetry between the scope of interests considered in the trolley scenarios (for which purpose, the account is taken only of the interests of the moral patients present in the particular situation),

well as an ML graduate student accusing the lab of being aligned with a lobbying strategy to manipulate AI ethics, is such a way as to avoid state regulation (Rodrigo Ochigame, 'The Invention of 'Ethical AI'. How Big Tech Manipulates Academia to Avoid Regulation' (*The Intercept*, 20 December 2019)).

[42] Awad et al. (n 7) 63.

[43] J Clement 'Global VPN Usage Reach 2018, by Region' (22 July 2019), https://www.statista.com/statistics/306955/vpn-proxy-server-use-worldwide-by-region/ (accessed 21 November 2019).

[44] Not only could the software not distinguish a criminal from a doctor nor produce a precise approximation of pedestrians' ages (particularly when faces are recorded from the profile or back), but it also often fails to recognize human beings, such as when Google's software identified a black person under the gorilla label.

[45] Sven Nyholm and Jilles Smids, 'The ethics of accident-algorithms for self-driving cars: an applied trolley problem?' (2016) 19(5) *Ethical Theory and Moral Practice* 1275.

and AV dilemmas, whose normative dimension requires the assessment of the interests of all people who may end up in such a situation. Secondly, the trolley dilemmas focus strictly on the agent's moral responsibility, whereas AV dilemmas also require the assessment of their legal liability, which significantly impacts decisions and constrains rights to action.[46] What is more, these dis-analogies 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,[47] 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.[48] Finally, it has also been remarked that the humoristic perception of the dilemma may alter respondents' decision-making process,[49] which is not only problematic for the *fat man case* (the 'less fit' criteria in the MM) in which two-thirds of the respondents were reportedly laughing, but also for the *bystander at the switch case*, in which a third of them were reportedly laughing.

3.2. The inner fallacy of computational social choice applied to ethical choices

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 that, 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 these and responsible for their 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,

[46]Allen Wood, 'Humanity as an End in Itself', in D Parfit and S Scheffler (eds), *On What Matters*, vol. 2, (Oxford University Press, 2011) 58, 74–5.

[47]Kathryn B Francis and others, 'Virtual Morality: Transitioning from Moral Judgment to Moral Action?' (2017) 12(1) *Plos One*.

[48]Dries H Bostyn, Sybren Sevenhant and Arne Roets, 'Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas' (2018) 29(7) *Psychological Science* 1084.

[49]Christopher W Bauman and others, 'Revisiting External Validity: Concerns About Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology' (2014) 8(9) *Social and Personality Psychology Compass* 536.

nor be held accountable for the consequences of their actions. There is thus, for now, insufficient ground to question the moral status of such algorithms without jeopardising the responsibility around their consequences, and also to refuse considering them as 'moral proxies'. Jason Millar 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;[50] the moral proxy of a moral agent is thus another *moral agent* making a *decision* in the best interests of the first one. Although it may look only semantic, this distinction is crucial because unlike Greene et al. who focus on 'hybrid collective decision-making systems'[51] with the purpose of improving communication between humans and robots for better collaboration, Conitzer et al. and Noothigattu et al. intend to create an autonomous system of moral decision-making, both considering such a system to 'make moral decisions'[52] or 'make ethical decisions'.[53] Granting algorithms the capacity to take moral decisions greatly jeopardises 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.

Secondly, 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'.[54] 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 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 to address these issues nor contradictors to challenge their answers. This methodological 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.[55] The importance of

[50] Jason Millar, 'Technology as Moral Proxy: Autonomy and Paternalism by Design' (2014) *Proceedings of the 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering* 128.
[51] Greene et al. (n 38) 4150.
[52] Conitzer et al. (n 39) 4831.
[53] Noothigattu et al. (n 8) 20.
[54] Conitzer et al. (n 39) 4834.
[55] Jean-François Bonnefon said that 'philosophers had the luxury not to solve the trolley dilemma. But today, we will not have to solve it but to find a solution with which we feel comfortable', from my translation of 'Les philosophes avaient le luxe de ne pas résoudre le dilemme du tramway. Mais il

finding the right decisions to fairly allocate harm in each dilemma situation is thus explicitly subordinated to the HMI.[56] 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 signalling and fear placebo.[57]

However, people often change their minds about moral choices, whose volatility is highly and negatively correlated to the degree of information and deliberate reasoning that they result from. Imagine a journalist asking people on the street about their perception of the ideal income tax rate, without informing them he was appointed by Congress to pilot a tax reform. Respondents may certainly give him a much lower rate than the present one. Not only are they abused by the journalist who is hiding from them his survey's goal, making such answers illegitimate to use, 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 favour of a higher tax rate had they been aware of the number of public services these taxes were funding and taken more time to respond, 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.

In response to this, Noothigattu et al. 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]',[58] which is consistent with the idea that 'idiosyncratic moral mistakes made by individual humans are washed out in the aggregate'.[59] However, such aggregation does not reduce noise but only normalises answers around an average social belief, one which presents no guarantee of approximating the

va falloir aujourd'hui non pas le résoudre, mais trouver une solution avec laquelle nous sommes confortables' Gregory Rozieres, 'Les voitures autonomes doivent-elles vous sacrifier pour sauver un enfant ou un chien ?' *The Huffington Post* (24 October 2018).

[56]'Every day the adoption of autonomous cars is delayed is another day that people will continue to lose their lives to the non-autonomous human drivers of yesterday' (Shariff et al. (n 20) 696).

[57]Shariff et al. (n 20), argue that convincing people to buy AVs implies making them feel both safe and virtuous. Such results can be achieved by establishing virtue signalling – as the possibility for AVs buyers to show off their virtuous consumption (i.e. buying AVs) to others via ostentatious signals – and by 'educat[ing] the public about the actual risks […] in a calculated way', offering them fear placebo – defined as 'high-visibility, low-cost gestures that do the most to assuage the public's fears without undermining the real benefits that autonomous vehicles might bring' (Ibid. 695).

[58]Noothigattu et al. (n 8) 20.

[59]Conitzer et al. (n 39) 4834.

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 prioritising 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. If we now define mistakes as marginal inconsistencies between respondent's replies, (sometimes prioritising saving the many over the few, and sometimes not), nothing could then be inferred from it, as the asymmetry of replies may just as much result from a lack of attention than they could reflect an unaccounted factor that the simplistic aspect of the MM fails to identify. Additionally, whatever the aggregation process selected, the VBS necessarily remains limited by Condorcet's paradox and Arrow's impossibility theorem.

In contrast, Greene et al. acknowledge that 'aggregating just preferences may lead to outcomes that do not follow any ethical principles or safety constraints', and then suggest fusing rather than aggregating values and preferences, combining hard constraints for basic ethical laws with only two levels of satisfaction (yes or no), and relaxed constraints for preferences, which can be satisfied at several levels.[60] Such an approach is preferable because it implies defining specific basic ethical laws that need to be human-generated and rationally justified, and conversely because it permits the inclusion of utilitarian principles of preference maximisation under hard constraints of inviolable deontologist rights. However, were a system to succeed in suggesting the closest approximation of a potential consensus within a social group, this nevertheless would not qualify it as a moral decision. Considering a population struggling with a two-alternative scenario, where 50% of the population is in favour of option 1 and 50% in favour of option 2, a unique decision-maker may end up satisfying 100% or 0% of the population, according to the perceived legitimacy of the procedure it chooses.

4. The instrumentalization of the ethical discourse

In this section, I refute the heart of the moral claim justifying the VBS, namely the lack of alternative solutions and the HMI, both justifying the authors' approach to solve AV dilemmas in the last resort. I then explain why the MM is not only fallacious, but also dangerous for society, elaborating on both its psychological impact and its distracting effect.

[60]Greene et al. (n 38) 4147.

4.1. Refutation of the 'highest moral imperative'

So far, I hope to have demonstrated 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, using illegitimate data which was not collected for the purpose of solving the AV dilemmas. However, there is one argument that 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.

In defense of their approach, Noothigattu et al. 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?'.[61] 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 John Rawls, given out of context and without any further comment regarding Rawls's theory of justice.[62] In addition, the paper written by researchers at Microsoft Research is clearly not ethics-oriented but business-oriented, while citizens and legislators, when assessing the AV moral issues, may adopt a more ethical-oriented approach:[63]

> In keeping with the motivation of fairness in online advertising, our approach will permit […] the vendor, as much freedom as possible, without knowledge of or trust in this party. This 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.

[61] Noothigattu et al. (n 8) 1.
[62] Cynthia Dwork and others, 'Fairness Through Awareness' (2011) *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* 214–26.
[63] Ibid. 1.

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, *ceteris paribus*. They however tend to agree on most of the other aspects of the dilemma,[64] 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 which enables people to 'agree to disagree' in modern democracies. When investigating the dilemma to build a common identity in a multicultural state, Charles Taylor recognises 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.[65] This is nothing else than the democratic way for people to 'agree to disagree' in modern democracies. Conversely, because individuals may agree to disagree for good reasons, they may also disagree to agree for bad ones. This is the issue VBS-like initiatives are exposed to. Whereas they are 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, *a priori*, be in favour 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 that is perceived to be 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 early deployment of AVs would be one of these. HMI's backers include officials, such as Mark Rosekin, and manufacturers, like Tesla's CEO Elon Musk, but also academics, including Bonnefon et al.:[66]

> manufacturers and regulators will need to accomplish three potentially incompatible objectives: being reasonably consistent, not causing public outrage, and

[64]Considering the steering driver case, the only distinction between Thomson and Foot's conclusions relates to the modality of the agent's moral commitment. The driver 'should' (moral obligation) turn the trolley for Foot, whereas he 'may' (moral permission) for Thomson (n 6) 206–7.

[65]Charles Taylor, 'Political Identity and the Problem of Democratic Exclusion' *ABC Religion and Ethics* (2016).

[66]Bonnefon et al. (n 26) 2.

> not discouraging buyers. Not discouraging buyers is a commercial necessity – but it is also in itself a moral imperative, given the social and safety benefits AVs provide over conventional cars.

Considering the size of the political and economic interests at stake behind AVs, together with the evidences of recklessness and outright negligence characterising the AV industry and self-driving tests,[67] it is not unreasonable to question the sincerity of such philanthropic ambitions. This, however, does not temper the need for a reasoned refutation of the HMI, which 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 authorise their marketing as early as AVs can reduce the net balance of total annual death by one,[68] to implement regulation with low liability for manufacturers in case of accident in order to avoid discouraging them,[69] and even to prohibit the use of NAVs when AVs become safer than them.[70]

Let us firstly agree on the fact that although a private company's employees may have personal moral obligations conflicting with their activity in the company – as recently illustrated by Google's internal oppositions regarding Maven, Jedi and Dragonfly projects[71] – the company itself is only legally bound by compliance to the law, and only ethically constrained by the provisions its shareholders have decided to include in its articles of association, specifically their social objects. Subsequently, unless Tesla, Waymo and Uber have explicitly included it in their statutes, they have no moral obligation to develop AVs to 'save' people who may die on the roads in the future.[72] Secondly, to be taken seriously, HMI's backers would also 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 unborn).[73] Much more significant and assured results in terms of the saving of life could indeed be reached by investing AVs' R&D budgets to feed the 821 million undernourished people worldwide.[74] An alternative solution for governments would also

[67] Heather Somerville and David Shepardson, 'Uber Car's 'Safety' Driver Streamed TV Show Before Fatal Crash: Police', *Reuters* (22 June 2018); Brian Merchant, 'The Deadly Recklessness of the Self-Driving Car Industry' *Gizmodo* (13 December 2018).

[68] Kalra and Groves (n 23).

[69] Hevelke and Nida-Rümelin (n 21); Bonnefon et al. (n 26).

[70] Sparrow and Howard (n 24).

[71] David Samuels, 'Is Big Tech Merging With Big Brother? Kinda Looks Like It' *Wired* (23 January 2019).

[72] It is one thing for a company to promote in its code of ethics the safety of its customers when using its products, and another to include the will to fight against road deaths in its statutes, orientating its activity toward this end.

[73] The HMI supporters usually refer to all the future lives that could be saved, including those of people not born yet.

[74] Food and Agriculture Organisation, *The State of Food Security and Nutrition in the World* (2018) United Nations, 2.

consist in enforcing regulation lowering the authorised speed limits to 25 km/h on all roads, what could even result in a fairer society,[75] rather than allowing an unprecedented regulation gap.[76]

Another argument against the claim that delaying AVs results in sacrificing lives[77] is given by Lin[78] who relates the issue to Derek Parfit's non-identity problem.[79] Although AVs may halve the number of deaths due to car accidents, Lin says the people who will still die on the roads 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 1000, resulting in changing the identity of many victims. The equivalence of deaths presupposed by utilitarians may then be challenged when considering the net average level of responsibility. Assuming that 94% of NAVs accidents are caused by human error[80] (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.

Finally, unlike private companies, governments do have a political objective to promote safety on roads. However, the fact that some people make dangerous use of NAVs is insufficient to support the paternalistic measure to oblige all of them to adopt AVs, prohibiting the use of NAVs.[81] By comparison, while the duty for a government to maximise the safety of its soldiers is arguably more stringent than that toward motorists, it is however insufficient to prevail over other ethical considerations as illustrated by the literature on lethal autonomous weapons systems. Therefore, 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 favourable

[75] Ivan Illich, *Energie et équité* (Seuil 1973).

[76] Joan Claybrook and Shaun Kildare, 'Autonomous Vehicles: No Driver … No Regulation?' (2018) 361 (6397) *Science* 36.

[77] Bauman and Youngblood (n 22): 'What should we do today so that over time autonomous vehicles become as safe as possible as quickly as possible without sacrificing lives to get there?'

[78] Patrick Lin, 'The Ethics of Saving Lives With Autonomous Cars is far Murkier Than You Think' *Wired* (30 July 2013).

[79] Derek Parfit, *Reasons and Persons* (Oxford University Press 1986).

[80] It is worth noting that this number is cited by the great majority of AV-related papers, including official reports from the European Commission, and used as a ground truth to justify the HMI. Despite its great influence, no concern has been raised about the fact that it comes from a study conducted by the NHTSA (which is openly engaged in favour of AVs' development), six years ago, based on a narrow sample of 5,470 crashes, all located in the U.S., and which did not necessarily imply fatalities Additional independent research should have been conducted to challenge these results.

[81] Jean-Baptiste Jeangène-Vilmer, 'Terminator Ethics: faut-il interdire les "robots tueurs"?' (2014) 4 *Politique étrangère* 151, 163.

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 – deriving from the foundation of their legitimacy rooted in the necessity to respect people's individual rights, especially in terms of equality and non-discrimination – cannot be subordinated to the opportunity offered by AVs to save a number of people's lives when engaged in a dangerous 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'.[82] 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.

4.2. The negative impacts of the MM on society

It is tempting to fall into 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, one psychological and one distractive, that prove it wrong.

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. are relevant to conduct life arbitrations. As demonstrated somewhere else,[83] not only are most of these criteria morally irrelevant (some of them even clearly unconstitutional), but the whole MM's characteristic-based approach is dangerous in itself, polarising 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. I had the opportunity to empirically observe these effects on my own students. While an overwhelming majority promptly raised their hands to arbitrate between men versus women or slim people versus less slim ones, most of them refused to answer when being asked who should be saved between a black and a white person, or a Muslim versus a Christian.[84]

[82] Federal Ministry of Transport and Digital Infrastructure (FMTDI), *Ethics Commission on Automated and Connected Driving* (2017) Report, 20.

[83] Hubert Etienne, 'A Practical Role-Based Approach to Solve Moral Dilemmas for Self-Driving Cars' (forthcoming).

[84] I do not think any experimental evidence is necessary here. For those who may disagree, I suggest that they should wonder whether the MM would have had the same greeting from newspapers and public opinion if its authors had based their scenarios on ethnicity, skin colour or religion, rather than on gender, weight and social status.

Second, a side-effect of the MM popularity was to distract from other first-order ethical issues. Some of them include preserving the integrity of embedded systems against hacking and threats of AV use for terrorist ends,[85] losing the possibility to react in critical situations (e.g. exceeding the authorised speed limit for medical emergencies, to escape impending aggression, or to run over a threatening gunman), and realising the importance for a country to have its own navigation systems for national independence. The impacts of AVs on the job market have also been raised and millions of workers may soon be undergoing career shifts considering that about 11% of American workers drive as part of their work.[86] I would now like to elaborate on two major issues which have comparatively received very little attention so far: the forthcoming prohibition of NAVs and the building of an AV-based mass surveillance system.

While two scenarios coexist regarding the deployment of AVs, it is clear that a mixed-traffic scenario could be nothing else than a transitory step towards an inevitable fully autonomous traffic, alongside the prohibition of NAVs as notably announced by Elon Musk.[87] Only such conditions would lead to the full extent of AVs' expected benefits, including the removal of redundant expensive signage, reducing the number of accidents, increasing the speed limit, closing the safety distance between cars, and improving traffic fluidity with intersection traffic management algorithmic regulators.[88] Such a prohibition may then arrive either from the law – governments would declare NAVs too dangerous,[89] justifying their removal for safety reasons similarly to how many cities are progressively banning diesel and even fuel cars for ecological reasons, and considering that the zero death objective claimed by both the US and EU authorities could only be achieved in a fully autonomous environment.[90] Or it could arrive through the market – insurance companies setting prohibitive prices for NAVs, while AV

[85]Lin presents a convincing argument against the probability of AVs being weaponized for terrorist goals, highlighting their cost-inefficiency when explosive drones can achieve similar results. Although it is right that AVs do not represent a substantial new opportunity for terrorist groups, he may however underestimate the weight of symbols for such organizations, as terror economy does not follow the same rules as the rational agents-based modern economy (Patrick Lin, 'Don't Fear the Car Bomb' *Bulletin of the Atomic Scientists* (17 August 2014)).

[86]David N Beede, Regina Powers and Cassandra Ingram, *The Employment Impact of Autonomous Vehicles* (US Department of Commerce, Economics and Statistics Administration, 2017).

[87]Stuart Dredge, 'Elon Musk: Self-Driving Cars Could Lead to Ban on Human Drivers' *The Guardian* (London, 18 March 2015).

[88]Tsz-Chiu Au and Peter Stone, 'Motion Planning Algorithms for Autonomous Intersection Management' (2010) *Proceedings of the 1st AAAI Conference on Bridging the Gaps Between Task and Motion Planning* 2–9.

[89]Sparrow and Howard (n 24) 209.

[90]Having progressively expelled old cars from Paris, the mayor recently decided to ban all diesel vehicles from circulating in the city as early as 2024, and all fuel cars by 2030. Such prohibitions for ecological reasons could absolutely be transposed to NAVs for safety motives, just as Bill Gates predicted (www.rtl.fr/actu/insolite/pour-bill-gates-conduire-sa-propre-voiture-sera-un-jour-illegal-7782354890).

insurance is included in their sale price.[91] The forthcoming prohibition of NAVs is problematic, as it will force people to significantly change the way they relate to their freedom of movement – 9% of Americans do not want to ride an AV because they enjoy the physical act of driving[92] – and put them in a situation of extreme surveillance.

AVs are equipped with an exhaustive range of sensors, including odometry, infrared and ultrasonic sensors, inertial and satellite navigation systems, as well as radars, lidars and cameras. The information they capture is relayed to the manufacturer's network as part of a continuous flow allowing the sharing of traffic information in real time. This raises major privacy issues (notably the images captured by internal and external cameras together with positioning information) and unprecedented concerns for individuals' surveillance. Even more so, would this data be made available to public authorities – as proposed by the European Commission:[93]

> as some of the data generated by vehicles may be of public interest, the Commission will consider the need to extend the right of public authorities to have access to more data. In particular, it will consider specifications under the Intelligent Transport Systems Directive regarding the access to data generated by vehicles to be shared with public authority.

AV data sharing with public entities is unavoidable, both because there will come a time when a common authority will be needed for traffic management purposes (e.g. intersection traffic management), and because all public transportation services expect an autonomous future. Therefore, the real danger does not relate to public authority access to AV data, but rather to the strength of the wall securing this data in the hands of an independent authority and preventing national security agencies from accessing it for surveillance purposes. There is not much doubt that AV sensors will be tied to the existing 176 million-strong public camera network using facial recognition to monitor the Chinese population. Even in France, often cited as one of the most protective countries regarding privacy and personal data, facial recognition experiments in public spaces for security purposes were recently authorised in Nice. The mayor, Christian Estrosi, deployed facial recognition systems in the city's public cameras network 'to track all comings and goings, on public transit, arteries, public places' of a list of individuals identified as potential threats for state security, arguing that 'we should put all the possible innovations at the service of our security'.[94]

[91] This option was already tested by Tesla in Asia (Danielle Muolo, 'Tesla Wants to Sell Future Cars With Insurance and Maintenance Included in the Price' *Business Insider France* (Paris, 23 February 2017)).

[92] Aaron Smith and Monica Anderson, 'Americans' Attitudes Toward Driverless Vehicles' (Pew Research Center, 2017).

[93] European Commission (n 4) 13.

[94] My translation of 'pouvoir suivre toutes les allées et venues, dans les transports en commun, dans les artères, dans les lieux publics, des individus en question' and 'Nous devons utiliser toutes les

Combining the exterior surveillance of the streets with the interior surveillance of the passengers through embedded vocal assistants, AVs may then become governments' eyes and ears.

They may ultimately become their hands, as a police interception tool to take control over fugitives' vehicles, as a discrete means to eliminate dissidents, driving them to unfrequented places to be assassinated without witnesses, or as weapons to stop public enemies. With regard to the latter, consider the following situation: Leo is leaving South Manhattan in his Tesla AV, making his way to the Sunday family brunch at his parents' place at Farmington, Connecticut. Meanwhile, the FBI has been notified that Alex, a young radicalised man whose track had been lost some days ago, is about to commit a terrorist attack on the Queensboro Bridge, exceptionally crowded because of the Marathon. The FBI has no information on Alex's position and no time for investigation; they call the New York central station of traffic management for help, which identifies Alex from Leo's Tesla cameras, only one street away from the bridge. On the FBI's order, Leo's Tesla suddenly accelerates, climbs on the sidewalk and fatally hits Alex. Without his consent, Leo's private car just became a national security weapon.

5. 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 century 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. Although computational approaches may not be abandoned, they should however 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 irresponsible non-human 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,[95] which at first glance depicts an orator discoursing in front of a chamber of representatives, but actually represents Cicero

innovations possibles au service de notre sécurité' (Agence France Presse, 'Nice va tester la reconnaissance faciale sur la voie publique' *Le Monde* (Paris, 18 February 2019)).

[95] www.media.mit.edu/projects/a-voting-based-system-for-ethical-decision-making/overview/.

denouncing the Catiline's plotting to the Senate and its dangers for the Roman Republic.[96]

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Hubert Etienne is a French philosopher conducting research in AI ethics and computational social sciences at Facebook AI Research and Ecole Normale Supérieure. He is a lecturer in Data Economics at HEC Paris, a lecturer in AI Ethics at Sciences Po, ESCP Europe and Ecole Polytechnique, as well as a research associate at Oxford University's Centre for Technology and Global Affairs.

ORCID

Hubert Etienne http://orcid.org/0000-0002-7884-1614

[96]*Cicerone denuncia Catilina*, fresco of Cesare Maccari, 1889, Roma, Palazzo Madama, Roma.