Article

The Immoral Machine

JOHN HARRIS

Abstract: In a recent paper in *Nature*[1] entitled *The Moral Machine Experiment,* Edmond Awad, *et al.* make a number of breathtakingly reckless assumptions, both about the decisionmaking capacities of current so-called "autonomous vehicles" and about the nature of morality and the law. Accepting their bizarre premise that the holy grail is to find out how to obtain cognizance of public morality and then program driverless vehicles accordingly, the following are the four steps to the Moral Machinists argument:

1) Find out what "public morality" will prefer to see happen.
2) On the basis of this discovery, claim both popular acceptance of the preferences and persuade would-be owners and manufacturers that the vehicles are programmed with the best solutions to any survival dilemmas they might face.
3) Citizen agreement thus characterized is then presumed to deliver moral license for the chosen preferences.
4) This yields "permission" to program vehicles to spare or condemn those outside the vehicles when their deaths will preserve vehicle and occupants.

This paper argues that the Moral Machine Experiment fails dramatically on all four counts.

Keywords: autonomous vehicles; driverless vehicles; public morality; the Moral Machine Experiment

In a recent paper in *Nature*[2] entitled *The Moral Machine Experiment,* Edmond Awad, *et al.* make a number of breathtakingly reckless assumptions, both about the decisionmaking capacities of current so-called autonomous vehicles, and about the nature of morality and the law. The assumptions and the habits of mind that they exhibit are of huge general interest, and of significance both for science and across the entire range of ethics in public affairs. In the first paragraph of their paper they say:

> *"Autonomous vehicles will need to decide how to divide up the risk between the different stakeholders on the road."*

It seems not to have occurred to the Moral Machinists that it is not open to the drivers of driverless cars, whether they are machines or humans, automatically to expose other innocent road users to injury or death when the alternative involves any risk to themselves or their machines.

The Moral Machine uses a decisionmaking methodology derived from the famous "Trolley Problem" invented by the Oxford philosopher Philippa Foot in 1967.[3] I was one of the group of (initially) Oxford philosophers who devised and popularized versions of this problem in the late '60s and early '70s of the last century.[4]

I am indebted to Giulia Cavaliere, John Coggon, Tomi Kushner and David Lawrence for many helpful discussions of these issues throughout this paper and to Philippa Foot for discussions between 48 and 50 years ago.

Cambridge Quarterly of Healthcare Ethics (2020), **29**, 71–79.

doi:10.1017/S096318011900080X

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

If a vehicle was really autonomous, instead of simply metaphorically autonomous,[5] it might do some deciding, but so-called autonomous vehicles are incapable of *deciding*; they will merely do some programmed *selecting* between alternatives. Better surely to call them "driverless vehicles" rather than "autonomous vehicles."

The Moral Machinists seem to believe that they have found a way of making themselves, and the rest of us, "cognizant of public morality," and that such knowledge will give them and us access to morally optimal, or at least morally acceptable, choices. This will, *in extremis*, so they think, minimize harm in proven morally acceptable ways. They also claim these ways to be somehow morally licensed by "public morality."

So…rather than doing what they claim "useless" ethicists do, the Moral Machinists have a cunning plan….

> *In other words, even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo. Any attempt to devise artificial intelligence ethics must be at least cognizant of public morality. (paragraph 3)*

We will return to the problematic nature of The Moral Machinists' account of public morality in a moment. But we should note that, in the real world, cognizance of, and respect for, the law is a much bigger hurdle for these cavalier experimenters. Many jurisdictions, including English Law, do not recognize a defense of necessity to charges of murder.[6] If, Dr Awad, for example, were, deliberately to drive his car into a bus queue, or even into an innocent jay walker, and caused death to avoid what he judged to be greater harm, he would, in many jurisdictions, rightly[7] find himself charged with murder or at the very least, with culpable homicide or manslaughter. So never mind "*artificial intelligence ethics*" needing to be "*at least cognizant of public morality*"; much more urgent is the necessity to be cognizant of the Law, a cognizance of which the Moral Machinists show no evidence whatsoever! Let's consider…

Two Landmark Legal Cases

In 1884, a landmark case in England established a precedent throughout the common law world that necessity is no defense to a charge of murder. In this case, two shipwrecked sailors, Dudley and Stephens, when a third survivor (the cabin boy Richard Parker) fell into a coma, decided to kill him for food to save their own lives and that of a fourth survivor.[8] Dudley and Stevens were convicted and sentenced to hang, but were reprieved and in fact served 6 months in prison.[9]

One hundred and sixteen years later in the year 2000, the questions of when and why might it be permissible to kill one human individual to save another, of how life and death choices between individuals can be justly made, were questions debated over many months in the United Kingdom, both in public and through the courts. These are the issues raised by the case of the so-called "Manchester conjoined twins" which captured the public imagination not only in the United Kingdom but internationally.

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

The drama turned essentially on the legitimacy of killing one of the twins so that the other might be saved, in conditions in which it was impossible to save both, and as to who should decide such an issue. I studied and wrote about[10] this case at the time, met the consultant surgeon involved, and followed the court proceedings and the appeal.

The twins were born on the 8th of August 2000, their bodies fused from the umbilicus to the sacrum,[11] and the lower ends of their spines and spinal cords also fused.

Jodie seemed neurologically normal, whereas Mary had a number of severe brain malformations and abnormal neurological responses.

An elective procedure to separate the twins involved an estimated mortality risk of around 6%. Any separation operation would lead to the death of Mary. It was believed that, although Jodie would have to undergo a series of operations through childhood to correct her congenital malformations, she would eventually be able to lead a substantially normal life if separated from Mary.

On the 18th of August, 10 days after the birth, the hospital initiated proceedings in the High Court under the Children Act 1989 seeking:

> A declaration that in the circumstances where (the children) cannot give valid consent and where (the parents) withhold their consent, it shall be lawful and in (the children's) best interests to (a) carry out such operative procedures not amounting to separation upon (Jodie and/or Mary), (b) perform an emergency separation procedure upon (Jodie and/or Mary) and/or (c) perform an elective separation procedure upon (Jodie and Mary).[12]

This declaration was granted on 25th August.[13] Both the parents and Official Solicitor acting on behalf of Mary appealed. On 22nd September, the Court of Appeal dismissed the appeal and upheld the declaration. On 6th November, the elective separation operation was performed, Mary died in an operating room, and, Jodie was expected to enjoy a relatively good quality of life with her family thereafter.[14]

This is what it takes in a mature democracy to decide just one such case. Such decisions involve a careful weighing of the evidence and of all the circumstances, the issue of consent and what, absent consent, can be imposed on innocent citizens.

The Moral Machinists purport to carry out this 'process' in advance, using superficial sampling in the form of an online "game," and believe this could yield results that will enable owners, manufacturers of, and passengers in, driverless cars to settle in advance the legal and ethical ramifications of any deaths resulting from the programming of the vehicles. This is naïveté of heroic proportions!

What is Public Morality?

Accepting for a moment, that the holy grail is to find out how to obtain cognizance of public morality and then program driverless vehicles accordingly, the following four steps are offered by the Moral Machinists as crucial:

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

1) Find out what "public morality" will prefer to see happen in a range of scenarios.
2) On the basis of this "discovery" claim both popular acceptance of the preferences and persuade would-be owners and manufacturers that the vehicles are to be programmed with the best and most acceptable solutions to any survival dilemmas they might face.
3) Citizen agreement, thus characterized, is then presumed to deliver moral license for the chosen preferences.
4) This yields "permission" to program vehicles to spare or condemn those outside the vehicles when their deaths will preserve vehicle and occupants.

Unfortunately, they fail on all counts. Here's why: The Moral Machinists do not demonstrate the first idea of what morality is, or might be, and hence what being "cognizant of public morality" might amount to.

Ronald Dworkin, one of the greatest Jurisprudential and Constitutional lawyers and philosophers of recent times, drew a distinction which is of crucial relevance here. In the context of a discussion of the famous debate between two other leading lawyers, Lord Patrick Devlin and H.L.A. Hart, concerning the enforcement of morality, Lord Devlin had made great play of the importance of respecting public opinion and community values. Dworkin's telling rebuke was: "What is shocking and wrong is not his [Lord Devlin's] idea that the community's morality counts but his idea of what counts as the community's morality."[15]

So-called autonomous vehicles do not solve moral dilemmas, if they did, if they only could, they would be 'persons,' properly so-called (super-intelligent AI persons), and would therefor necessarily have rights, interests, duties, and votes, like the rest of us persons. Much more important, their 'lives'—we should more appropriately say their 'existences,' would matter, and would count equally with that of humans.[16] Alas, they would, in the world of the Moral Machinists, have preferences and make decisions without solving the dilemmas that makes those preferences moral.

"Exactly as not just any judgment about things in which science is interested is 'scientific'"[17] or a part of science, so not just any judgments about things with which morality is concerned are moral judgments.[18] The solving of a moral dilemma involves much, much more than having a preference for one possible outcome of a moral dilemma, just as the resolution of a scientific problem requires much more than simply opting for (stipulating) a particular solution! It has to show how the circumstances which make it a moral dilemma, have been weighed carefully one against another, and morally persuasive reasons, facts and/or justifications found for having a moral preference for one outcome rather than another. Majorities are not necessarily right; neither science nor ethics is produced by casting votes for particular 'answers'; happy though such a possibility might seem to some! The Moral Machinists are proposing the moral equivalent of deciding whether the world is flat by finding out what people would prefer the answer to be.[19]

Tossing a coin, to be sure, selects an outcome, but not for moral reasons.[20] Neither coin tossing, nor algorithm 'obedience,' nor the methods described in the Moral Machine Experiment paper, are methods or processes of moral deliberation! Nor do they seem to have resulted from much deliberation of any kind.

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

The Moral Machinists work is "useless," as they claim that of ethicists would be, in the paragraph quoted above. They have not discovered whether citizens and corporations are, or are not, prepared to accept the moral and legal consequences of the decisions of the autonomous vehicles they might either own, or travel in, or have manufactured. Among which will be the consequences of deliberately causing the deaths of innocent bystanders and other road users, consequences which involve much more than operationalizing a principle of minimizing the *immediate* harm of directing a vehicle. Proper consideration of the possible harms of the preferences delivered by the Moral Machinists must surely include their effect, for example, on due process, the principle "that no person should be condemned unheard," on constitutional protections for freedom of the individual, and on justice, including criminal justice. All these are considerations of which the authors of The Moral Machine, and their experimental subjects, exhibit no awareness whatsoever!

Having a preference for killing (or sparing) one person rather than another (even a preference shared with thousands or even millions) doesn't make it moral. The preferences of a certain "Bohemian corporal,"[21] it is salutary to remember, came to be shared by millions.

Laws Arrived at Democratically, and Over Time, are One Indicator.

The helpful drawing provided by the Moral Machinists (1. b. in their paper) for their subjects, which shows a driverless car with the option of mowing down three old people on a crossing or driving into a solid wall, is a proverbial "no brainer." A human driver should surely steer for the solid barrier relying on the cars collision technology, crumple zones, air bags etc., to keep the occupants safe, or, if not safe, then at least alive. They should hope for the best, rather than self interestedly arguing to themselves that it doesn't matter that they save themselves because the 'old codgers' on the crossing have ignored a 'do not cross' sign, and so deserve everything they get!

In any event, there could be no certainty that the passengers in the driverless vehicle would be killed or even injured; and they certainly would be at less risk in their vehicle, than the unprotected old people on the crossing. I can see no rational, nor any moral basis in this example for a choice (human or machine) to do anything but try to avoid the old people.

Surely the right thing to do is to design better safety cages in driverless cars, rather than programming the AI controlling them to select convenient, cheap, lazy and 'soft' targets derived from a computer game, for sacrifice.

We have no space here to talk further about the totally trivializing setting of the questions, in which the moral machinists are inviting respondents to make their life or death selections, as simple expressions of what they "prefer."

We also need to recall that the law, both civil and criminal, is also an expression, if incomplete, of public morality; and if not necessarily more reliable, it is at least more soberly arrived at, than the snap expression of preferences relied on by the Moral Machinists in their "experiment." The Moral Machinists should have been aware that they also need parliaments, legislators, the courts, human rights conventions and many others (not simply a crude *vox pop*) not only to be cognizant of their lethal plans, but have debated them, consulted about them, and legislated to accommodate them, possibly in contravention of considerations to which we will now return.

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

Capital Punishment for Jaywalkers?

The Moral Machinists ask (paragraph 4):

> *[W]hether people prefer to spare the young rather that the elderly or whether they prefer to spare pedestrians who cross legally rather than pedestrians who jaywalk.*

Any decent citizens would immediately ask: What gives you the right to condemn to death innocent people, even ones so wicked as jay-walkers,[22] without examination of the question as to whether anyone has acted illegally, a due process, which commands public respect, which has heard their defense and convicted them, especially in a jurisdiction (unlike the U.K.) which has not abolished the death penalty?

The Moral Machinists seem unaware that most such matters of life or death have not been left to the bare preferences of people who have shown no evidence of deliberation. All juries for example, are cautioned about their responsibilities, required to hear and pay attention to evidence and argument from both sides, (prosecution and defense and also from the judge) and *deliberate*. I find no evidence of consideration in the Moral Machine Paper, nor any evidence of deliberation by those consulted. "Public morality," as they crudely and mistakenly understand it, requires only ill informed, unconsidered preferences, given instantly and thoughtlessly, as if playing a computer game!

Due Process

Rather, it is the case that public morality expressed through laws, civil and human rights conventions, and in many other publicly accessible ways in most civilized societies, has not simply been left to individual 'preference,' let alone to popular preferences or to popular prejudice, nor yet to psychologists, nor even to philosophers! It has evolved over a lengthy period, often painfully; informed by history, art, literature, culture, personal experience, and much more. It cannot just be bolted on by a Q and A: 'Do you prefer picture A or picture B?' There is, for example, a difference between a 'preference' and a 'prejudice,'[23] let alone a moral judgment and a prejudice. Consideration of that difference is nowhere evident in the prejudices (or is it simply the preferences?) of the Moral Machinists or their experimental subjects.

If the answers to the questions chosen by the Moral Machinists are remotely relevant, why are not answers to the questions as to whether or not citizens prefer to spare white people, or women, or priests, or those in military uniforms, not equally crucial? The answer is that these machinist questions have been put through a tendentious 'moral' filter, not a filter of data about popular morality or knowledge of ethical reasoning, but a question-begging filter of their own devising. Would the authors accept that if asked they should accept an overwhelming preference for "sparing" only white men (or only black women for that matter) as a part of public morality which required respect, let alone implementation by a machine, without any checks or balances? And what about their apparent embracing of the idea of summary 'justice' meted out without any due process on innocent victims? What does/did that tell the respondents about the moral seriousness, let alone the credentials of the questions and the questioners? What does it tell us

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

about the moral seriousness and the scientific credentials of the experimenters and of the journal which published their paper?

The idea that it might be open to individual citizens or corporations to decide who shall be “spared” and who condemned to death, and that this might be a matter of mere individual “preference,” made on the basis of the sorts of sampling described in their paper, whether of vehicle, or owner, or vehicle programmer, or population sampling, is outrageous in the extreme. The question “do you prefer to spare jay walkers or innocent passers by” beggars belief. Why not “do you prefer to spare Jews or gypsies, migrants or citizens, politicians or nurses, sports stars or vagrants?” It is an invitation to approve the summary execution of jaywalkers for something which, even if it constitutes a minor misdemeanor, is not something which the so-called jaywalkers have been either charged with or of which they have been convicted.

It is not, in most jurisdictions, simply up to individuals (whether those individuals are people or even driverless vehicles), to take the law into their own hands and refuse to “spare,” that is, to ‘deliberately execute,’ an innocent fellow individual for the ’crime’ of jaywalking, particularly in circumstances in which the self interest of the vehicle and passengers is obviously paramount.

Oates Law

I have a modest proposal[24] to make on this subject. No one should deliberately kill the innocent without an excuse of overwhelming, plausible, and judicially approved necessity, as in the Manchester conjoined twins case. If a vehicle is directable, but out of control, the driver should not deliberately kill others rather than put herself[25]/itself, its passengers and vehicle at risk. If we are to program driverless vehicles, the road would be much safer if the following was the first law of vehicle robotics. We may call it “Oates Law”.

On 1st November 1911 Captains Scott and Oates, and 14 other members of Robert Falcon Scott’s Antarctic expedition, set off from their Cape Evans base camp for the South Pole. “On 15 March, Oates told his companions that he could not go on and proposed that they leave him in his sleeping-bag, which they refused to do. He managed a few more miles that day but his condition worsened that night… On the morning of 16 March (or possibly 15 March – Scott was unsure) Oates walked out of the tent into a… blizzard and to his death. Scott wrote in his diary: “We knew that poor Oates was walking to his death, but though we tried to dissuade him, we knew it was the act of a brave man and an English gentleman.” According to Scott’s diary, as Oates left the tent he said, "I am just going outside and may be some time…" Oates most likely died on March 17th.[26] All of Scott’s party eventually died on this journey.

Oates legendary self-sacrifice to try to save the lives of his colleagues in the face of diminishing supplies of food and the need to make enough speed to reach further supplies, sets a moral example that might daunt most humans, let alone driverless cars. What we might expect of an autonomous vehicle, but probably not perhaps, simply of a driverless one is a question for another occasion?[27]

Whatever its priorities, it will surely be conscious that its solution to the problem will ‘say’ something about what sort of creature it is, and will influence, as does all decisionmaking, the sort of creature it will from thenceforth be, both in its own ‘mind’ and in the minds of others. But which ‘other minds’ (if any) will/

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

should it care about most? And of course, an analogous consciousness of what advocating or acting on the principles of the Moral Machine will say about the morality of those who have seriously proposed such a scheme for deciding how, ethically, we might program driverless cars.

Perhaps the most obvious answer to the spurious questions put by the highly amoral and indeed immoral, Moral Machine Experiment, is that driverless vehicles should always risk themselves and their occupants, relying on the safety built into the structure of their vehicles, rather than choose between different groups of innocent bystanders, not least because of the obviously corrupting self-interest involved. And all bystanders must be presumed innocent unless due process has found them guilty, and even then, it is not up to men or machines to devise their own additional penalties. True, not many people would want to ride in, or own, a driverless car with these priorities. But perhaps that is for the best; until, that is, we have real Autonomous Vehicles who can take their own moral responsibilities seriously, and who have had a proper education, not least in law and ethics.

Notes

1. https://www.nature.com/articles/s41586-018-0637-6/ (last accessed 20 July 2019).
2. https://www.nature.com/articles/s41586-018-0637-6/ (last accessed 20 July 2019).
3. Philippa F. The problem of abortion and the doctrine of the double effect in *Virtues and Vices*, Oxford: Basil Blackwell; 1978. (Originally appeared in the *Oxford Review* 1967;5.)
4. Harris J. The survival lottery. *Philosophy* 1975;50:81–8; available at https://en.wikipedia.org/wiki/The_survival_lottery (last accessed 20 July, 2019).
5. Harris J. Who owns my autonomous vehicle: Ethics and responsibility in artificial and human intelligence. *Cambridge Quarterly of Healthcare Ethics* 2018;27(4):500–9.
6. *Re A (Children)* [2000] EWCA Civ 254, [2000] 4 All ER 961. This case did in fact recognize necessity as a defense to what would otherwise have been an unlawful killing, but noted that the judgment applied to this case alone and did not set a precedent. See also Harris J. Human beings, persons and conjoined twins: An ethical analysis of the judgement in *Re A. Medical Law Review* 2002;9(3):221–36. See also *R v. Dudley and Stephens* (1884) 14 QBD 273 DC discussed in more detail below.
7. Harris J. *How to be good*. Oxford: Oxford University Press; 2016, and Harris J. *The value of life*. London: Routledge; 1985.
8. *R v. Dudley and Stephens* (1884) 14 QBD 273 DC.
9. https://en.wikipedia.org/wiki/R_v_Dudley_and_Stephens.
10. Harris J. Human beings, persons and conjoined twins: An ethical analysis of the judgement in *Re A. Medical Law Review* 2002;9(3):221–36.
11. The medical facts related here were agreed upon by several teams of doctors, including specialists from other hospitals called in as experts by the Court of Appeal. See also Note 6, Re A 2000, at 972, *per* Ward LJ.
12. Note 6, Re A 2000, at 987.
13. *Central Manchester Healthcare Trust v. Mr and Mrs A and A Child* (unreported).
14. Bunyan N. “Bright and alert” Jodie makes rapid progress. *Daily Telegraph Online*; available at http://www.lineone.net/telegraph/2000/12/16/news/bright_45.html, posted 16 Dec 2000 (last accessed 20 July 2019).
15. Dworkin R. *Taking rights seriously*. London: Duckworth; 1977, at 255.
16. See note 7, Harris 2016, chapters 1, 2, and 9.
17. See note 7, Harris 2016.
18. See note 7, Harris 2016, at 143.
19. See note 7, Harris 2016, chapter 2.
20. Unless there is no reason to prefer one outcome to another. Tossing a coin does not decide the better outcome, but does deliver an outcome.
21. https://history.stackexchange.com/questions/38799/why-did-paul-von-hindenburg-wrongly-call-adolf-hitler-bohemian-corporal (last accessed 20 July 2019).
22. Irony alert!

https://doi.org/10.1017/S096318011900080X Published online by Cambridge University Press

23. There is a considerable literature on this distinction. For an easy entry into that literature see Harris J. What it's like to be good. *Cambridge Quarterly of Healthcare Ethics* 2012;21(3):293–305.
24. Following Swift J. https://en.wikipedia.org/wiki/A_Modest_Proposal (last accessed 20 July 2019).
25. Perhaps cars, like ships, are conventionally female?
26. https://en.wikipedia.org/wiki/Lawrence_Oates (last accessed 24 July 2019).
27. See note 7, Harris 2016.