Philosophy and the Politics of Moral Machines

Paul Dumouchel*

Abstract

A recent large-scale survey, "The Moral Machine experiment" (2018) aggregated 39.61 million decisions across 233 countries and territories reflecting people's preferences as to who should be spared in fatal moral dilemmas involving autonomous road vehicles. The experiment collected 'big data' to reach conclusions concerning the moral rules that should be implemented in these vehicles. In this paper, first I question the philosophical presuppositions of the experiment, arguing that it has very little to do with ethics or moral norms, but essentially constitutes a market survey concerning the social acceptance of a dangerous technology. Then, I criticize the myth of moral machines and the illusion that abandoning to automated systems the power to 'autonomously' take lethal 'decisions' is a radically new phenomenon. Finally, I suggest a different solution to the difficulties addressed by the Moral Machine experiment and make political and legal suggestions concerning policy towards 'autonomous road vehicles'.

Key words : self-driving vehicles, moral preferences, moral machines, autonomy, responsibility, artificial intelligence, ethics

* Professor, Ritsumeikan University

〈 contents 〉

1. Introduction

What is a moral machine? What could be considered as one? What conditions are necessary in order for a machine to be moral, or immoral? Is it sufficient for a machine's action in the world to have (at time) consequences which have moral import in order for it to be a 'moral machine'? If the answer is 'yes', then it seems that just about any machine will be a moral machine, more precisely any artifact, machine or not, a medication, a knife, or a printing press will be moral. Clearly by 'moral machine' we mean more than that, but how exactly can this further demand be cashed out? Generally, the basic, necessary and perhaps sufficient requirement in order for a machine to be moral is for it to be autonomous. That is to say, to be in some way 'responsible' for its action in the world, the machine or artificial system should, so to speak, be 'in charge of its own action'. The problem is that autonomy is a concept which is hard to define and that means different things in different domains.

In AI, robotics, information science and the sciences of artificial systems, disciplines which produce autonomous artificial agents and machines, autonomy is defined as the ability of a system to adapt by itself to changes in its environment. It is thus defined relative to an environment, but as Lucy Suchman reminds us at the beginning of her classic *Human-Machine Reconfiguration*(L. Suchman, 2007). the environment in which a machine or artificial agent acts is quite different from that in which humans live. It is significantly poorer and more limited. In consequence, what is autonomous for a machine may be quite obviously and trivially determined from a human point of view. The second difficulty is that what "acting in a way that is adapted to the environment" means in the context of artificial systems, unlike what is the case for natural systems, primarily depends on what the system's designer or programmer want its to do. To be adapted here, is not a simple relation between the agent and its environment. It is also relative to a norm that is imposed from the outside by the systems creator or programmer. The question of the 'morality' of the machine then also needs to take into account that norm and its ethical value.

The issue is further made difficult by the fact 'moral autonomy' can be understood in a variety of ways. Here are three very different, yet, I think, serious candidates. First, to be morally autonomous in the Kantian sense is for an agent to give to him or herself, his or her own (moral) law. Which means to treat every person (including the agent) as an end in itself, rather than merely a means to an end. Alternatively an agent acts morally in the Kantian sense if the maxim of her action could be transformed into a universal law of nature. Second, an agent is morally autonomous to the extent that moral obligations to which the agent is subject, unlike laws of nature bind the agent but do not determine his actions. Or, to say it in another way, one's action is only morally autonomous if the agent could have acted otherwise, that is, immorally.[1)](M. Hildebrandt. 2015. p.296) Finally, a morally autonomous person is one who can recognize that another person has a legitimate claim, even in the absence of a pre-existing moral norms that justify that claim. This last formulation is close to Sen's idea of 'against injustice' and to Bergson's concept of 'open moral'. All three views of moral autonomy are different, but they are closely related in that all presuppose that a moral agent can in some way take distance from the rules that guide his or her action. Something that is particularly difficult to do for a machine, but where resides the whole issue of its possible moral autonomy.

Given these difficulties, answering the questions : "what is a moral machine?" or "can a machine be moral?" is far from evident. In order to begin addressing these questions, I will look at a recent inquiry on moral machines carried out on line, called the "Moral Machine Experiment".

2. The Moral Machine Experiment

A large scale research named "The Moral Machine experiment" (2018) whose results were recently published in the journal *Nature* aggregated millions of decisions across 233 countries, dependencies and territories reflecting, the authors argue, people's preferences as to who should be spared and who should be sacrificed, that is who should live and who should die, in fatal accidents involving self-driving cars.(E. Awad et al., 2018) The experiment was designed on the basis of a classic problem in ethics, named the trolley dilemma.[2] It used a free

[1] Note that a similar requirement applies to law, as Mireille Hildebrandt, reminds us. What distinguishes law from technological normativity is that it can be resisted.

access on line 'serious game', where players chose in different scenarios involving a self-driving car between killing, say, five passengers who are executives or three pedestrians who are homeless vagrants, or one baby and a cat over an elderly car passenger. As time went by the game allowed researchers to collect 'big data', that is, 39.61 million decisions involving nine factors that may influence people's choices : sparing humans vs pets; staying on course vs swerving; sparing more vs less lives; sparing men vs women; sparing the young vs the elderly; the fit vs the less fit; those who obey the law vs those who jaywalk; persons of higher vs those of lower social status. From the results obtained the authors draw recommendations concerning, the 'moral rules' which should be implemented in self-driving road vehicles and how such autonomous vehicles should be regulated by policy makers (2018 : 60).

The rationale for this experiment/survey is, say the authors, that we cannot avoid creating 'moral machines' as we will soon be faced with a radically new situations where autonomous vehicles will be called upon to distribute costs and benefits between stakeholders. The compromises required by this distribution, they argue, fall within the domain of ethics and we need to agree upon the rules that will guide the machines' choices. The goal of the experiment was to 'bring out' rather than to bring about this agreement. What I mean by 'bring out' rather than 'bring about' is that the objective was not to open a discussion, as would be the case in a deliberative democracy model, on the moral principles that should regulate the use of autonomous

[2] In the trolley dilemma an operator can either save five persons and kill one or vice versa save one and kill five by changing a switch to deviate an out of control trolley. The problem is : what is the moral thing to do? How do you choose? It was originally published by Foot, Ph. (1967). The Problem of Abortion and the Doctrine of Double Effect, *Oxford Review 5*.

vehicles, but to demonstrate *that there already is an agreement* among people's 'moral preferences' concerning the choices that autonomous vehicles should make in particularly difficult cases; an agreement which, according to them, their experiment/survey reveals.[3] The results showed that there is global preference for saving humans over pets, for saving more rather than less lives, for saving the young rather than the elderly (2018 : 61). Yet, whether these results reveal that there is an agreement as to which rules should be implemented in the machines and how they should be regulated is, I will argue, an uncertain claim, that is more difficult to establish.

These findings, the three common preferences can hardly be described as revolutionary or unexpected results. They seem to confirm most people's intuitions, or prejudices, concerning these issues. As we will soon see, this is not surprising given the design of the experiment. However, strong of the authority of these 40 million decisions[4] our authors argue that manufacturers of autonomous vehicles and policy makers need to take these results into account, need to implement in self-driving vehicles rules that will respect these preferences. Should they really?

The moral machine experiment is designed as a third person (still) video game. That is when you play the game you are neither in the car as a passenger, nor a pedestrian crossing the street, but an external observer who, as the website reminds you, is in no danger whatsoever. There is no character or avatar in the game that corresponds to you

[3] This raises the issue of the conditions under which statistical aggregation can be considered equivalent to an agreement.

[4] One question is that, as we do not know how many times an individual player came back, and thus how many games he or she may have played, 40 million represents the number of decisions, but it is difficult to judge exactly how many persons actually participated or if all entries should be weighed equally.

— the player — who suffers the consequences of your decisions. This is a game at which you cannot lose (or win for that matter). As such it corresponds to the kind of limited environment in which artificial agents are modelled to act. One in which the agent has no interest whatsoever. It is therefore unclear that a person placed is a similar dilemma would act according to her 'stated preferences', that is to those he or she expressed in the game as a third party.[5] Further, given that, as the authors remind us, drivers involved in such situation many times do not recognize that they are faced with a moral dilemma, (2018 : 59) we have little reason to assume that their 'revealed preferences', that is, those they would show to be theirs through their action (should they have to act) would coincide with their 'stated moral preferences'. The game is characterized by perfect knowledge which does not in any way correspond to the situation in which drivers have to take their decisions, nor in fact to the situation in which people usually make moral choices.

Therefore, the 40 million 'moral preferences' expressed here cannot be understood to be what would guide agents' action in similar circumstances or even the rules they would prefer should guide the choices of the vehicle should they happen to be a passenger. Rather they corresponds to the opinions of third parties concerning what 'autonomous' vehicles should do in situation involving harm or death of one or more persons, persons which they are not. Such preferences are only 'moral' in the sense that they pertain to issues that have moral import (life or death), but not in the sense they move agents, artificial

[5] In fact, there is good evidence that people prefer different rules if they are passengers rather that third party observers. See S. Guillebeault. (2019). *Le Bon, La Brute et le Truand. Ou comment l'intelligence artificielle transforme nos vie.* Montréal : Druide, pp. 58-62.

or natural, to act morally rather than immorally. To put it otherwise, radically immoral or unethical preferences are nonetheless 'moral preferences' in this sense. There is no reason therefore why such preferences should be considered morally binding, nor is it clear how morally relevant they are.[6]

It follows that this survey only has an indirect and tenuous relationship to ethics. Centrally it concerns something completely different which is why its results are thought to be relevant to manufacturers and policy makers. What the survey speaks to are the conditions of acceptance of a dangerous new technology. One which we are repeatedly told is unavoidable, is already upon us, and whose benefit cannot be forfeited. As the authors write : "For consumers to switch from traditional human driven cars to autonomous vehicles and for the wider public to accept the proliferation of artificial intelligence driven vehicles on their roads, both groups will need to understand the origins of the ethical principles that are programmed into these vehicles." (2018 : 59) The moral machine experiment has little to do with ethics, it is essentially a market survey that inquires into the conditions that will render acceptable to the general public a dangerous new technology : self-driving road vehicles.

It should be remembered in this context that the biggest market for autonomous vehicles is likely to be the trucking industry. Drones are already being used by some delivery services, but replacing drivers by autonomous vehicles for long distance haul and local transport will represent huge economies in terms of salaries and important gains in time and efficiency : autonomous trucks do not have to sleep, they do

[6] Actually, it may be argued that nearly all preferences are 'moral' in that sense, since there are very few issues, choices or decisions which do not have any, more of less important, moral consequences.

not have to eat and there is no law limiting the number of hours they can drive without resting. Furthermore, they neither have union nor retirement plans. Given the economic forces that will push for the adoption of this technology, it is therefore clear, in view of the probable proliferation of large self-driven trucks speeding down highways, that we are all interested as second parties, and as drivers/passengers, in the way these ‘autonomous’ vehicles will react in situations where they may cause harm and in the issue of responsibility should harm result.

Under this interpretation, as a survey concerning the introduction of a dangerous new technology the experiment’s conclusions are weakened by the fact that the majority of players were between the ages of 20 and 30.[7] Even if it may be argued that this age group corresponds to most of those who will acquire such vehicles either on a commercial or personal basis in the future, it does not reflect the age pyramid of societies where this technology is first likely to be adopted. Given this, it is not clear that the results correspond to the majority opinion of those who will be exposed as second parties (other drivers, passengers, users and pedestrians), rather than owners, to the dangerous consequences of this technological change. Nor does the Moral Machine experiment seriously inquire into what these consequences may be. It is confined to well trodden moral dilemmas corresponding to admittedly rare and unrealistic situations, but claims that this lack of realism is irrelevant as the point is to find agreement concerning moral principles! (2018 : 59) However, as Mathias Scheutz recently argued, the consequences of introducing autonomous vehicles may turn out to be surprising and raise unexpected new questions of moral and legal

[7] In fact a fair number may even have been under the legal driving age and more participants were attending high school than college.

responsibility. For example, “lack of coordination with humans through other channels (such as gestures or eye contact to indicate intent) could make autonomous cars genuine road hazards when injected into a system of human drivers, ultimately leading to destabilizing emergent properties of our whole traffic system when enough autonomous cars are present.”(M. Scheutz, 2017) How should such issues be addressed?

These limitations suggests that the experiment was not actually intended as a market survey, even though in the end that is what it turned out to be. Therefore, my analysis should not be misinterpreted as trying to denounce some hidden conspiracy, the attempt to disguise a market survey as an ethical inquiry. Rather the question we need to raise is why is was this inquiry into the ethical principles that should regulate the action of autonomous vehicles essentially construed as a market survey? Why is morality essentially viewed as a question of opinion and preferences? What allows this confusion of genres and encourages us to think that ‘moral preferences’ are necessarily moral?

3. Responsibility and Ethical Machines

To the extent that it claims to be about ethics, it is significant that in the Moral Machine Experiment the question of responsibility is never raised. This is rather surprising given that we should expect this question to be central for both manufacturers and policy makers to whom the experiment’s results are addressed as recommendations.8) Whether it is in the case of a dilemmas where harm cannot be avoided

8) A quick word search reveals no occurrence of the word ‘responsibility’ in the text or the annexes, while ‘responsible’ occurs only once in the title of one of the references.

for everyone, or in simpler and most frequent cases where avoidable harm takes place, the question of responsibility both moral and legal inevitably arises. The absence of the issue of responsibility in the Moral Machine Experiment comes, I believe, from the fact that in the mind of the researchers who designed this experiment the question had already been resolved. It was viewed as settled long before the experiment began.

Responsibility in their mind had already been shifted to the machine, for that displacement of responsibility to the machine is in fact what the experiment is all about. The reason why we need moral principles to guide the decisions of self-driven vehicles is because want autonomous cars to act ethically, responsibly. This shifting of responsibility from humans onto machines is common in reflections on moral machines and in robot ethics, as are two other ideas which are also shared by our authors. The first is that the displacement of responsibility from human to artificial agents is inevitable. The second is that the ability to 'decide' upon moral issues by machines or artificial systems is a radically new phenomenon. The two ideas are closely related. It is because the ability to act upon moral issues by artificial system constitutes an absolute novelty, that we are, so to speak, doomed to invent 'moral machines'. An innovation that is proportionate to the radical transformation that makes it necessary. The only possible response to a never before encountered transformation.

Let us therefore start by the second of these two ideas. "Never before in the history of humanity", claim our authors, "have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real time supervision. We are going to cross that bridge any time now…" (2018 : 63). Is that really the case?

Think, for example, of a land mine which ‘autonomously decides’ in a fraction of a second, without real time supervision, that whoever stepped on the trigger shall die.[9] Going back in time, think of a foot trap with spikes at the bottom of the pit to receive the falling body. Or think, as a different type of example, of fire protection engineering defined as the use of engineering principles to protect people, property and the environment from the harmful effects of fire and smoke. It is clear that all three functions cannot be maximized simultaneously. Any solution that is optimal relative to one objective, say protect property, is unlikely to be so relative to the other two. It follows therefore that compromises and trade offs must be made. These will be materialized in special devices, for example fire doors that automatically lock and close, or simply in the architectural layout of the building. In the event of a fire if any death occurs, we will of course invoke bad luck, but it remains that different decisions to privilege this rather than that function in certain circumstances may have lethal consequences and that these decisions are inscribed at various points in the design of the building, an object which most architects and engineers tend to consider as a machine.

It is quite clear that, for a very long-time already, and for many different reasons, sometimes simply because we are unable to do otherwise, humans have repeatedly shifted the decision of who will live and who will die unto artificial systems. What they have not done, however, is to simultaneously transfer to these systems themselves the responsibility for these decisions, though they have often argued that in

[9] The objection that this case is different because landmines are used in the context of warfare where the issue of the morality of killing has already been resolved is not very convincing given that most casualties of land mines are civilians and happen long after the conflict is over.

such cases nobody is responsible. Dislocating lethal decisions from humans to artificial systems is something that had already been done in very simple and ancient societies and has continuously been done since. What is new here is the endeavour to hand the responsibility over to the machine. Yet, if the situation is not as unprecedented as it is claimed to be, if for centuries we have at times allowed artificial systems to decide who should die and who would live, why do we need ethical machines now? Why do we need artificial agents that can take responsibility for their action? Or can they?

4. Intelligent and Ethical Machines

It may be argued of course that these ancient machines were not very smart. That is certainly true. Yet even a landmine can be claimed to have a minimal form of autonomy. It perceives some aspects of its environment and reacts to it adaptively, for, as mentioned before, in the context of artificial systems adaptation refers primarily to what we want the system to do. However, a landmine certainly does not have any moral autonomy. The fact that a system is more or less 'intelligent' is without doubt relevant here, but should greater intelligence be a sufficient reason to declare the machine responsible? Can a machine's greater intelligence make it an ethical machine? If so what level of intelligence is necessary? Replace landmines then, say, by extremely smart drones that can recognize enemy combatants, distinguish them from civilians, anticipate collateral damages, and decide autonomously whether to fire or not, is this level of complexity and autonomy sufficient to make these drones morally responsible? And if it is, what does the machine’s moral responsibility exactly mean? Let us suppose that we agree that our smart drone is responsible and that in one incident it kills innocent bystanders. Are those who deployed that ‘autonomous systems’ thereby relieved of responsibility, given that it is the drone, not them, which did it? Within a military context, that conclusion seems unlikely.

In fact, a major argument in favour of battlefield robots, according to Ronald Arkins, robot specialist and military engineer, is their *inability* to decide by themselves — that is, autonomously — “the moral implications of the use of lethal force”. This restriction of the machines’ moral autonomy is, he argues, (paradoxically) done by transforming battlefield robots into ‘thical robots’. That is, by programming them to apply the rules concerning the use of lethal force "that have been previously derived by humanity as prescribed by the [laws of war] and the [rules of engagements].(Ronald C. Arkins, 2009. pp. 116-117) In other words, the great advantage of ‘ethical military robots’, is that *they are not morally autonomous.* Why is this such an advantage? Simply because morally autonomous agents will inevitably sometimes be insubordinate, question the orders they receive, for ethical or for other reasons. Arkins’s battlefield robot project aims in reality to kill two birds with one stone. First, to provide commanders with a troop that is perfectly obedient. Battlefield robots cannot but comply with the orders they receive. Second, programming them to follow the laws of war and the rules of engagement (of the American army), Arkins claims, ensures that these machines, unlike human soldiers, will follow these rules without fault. They will never engage in atrocities because of their emotions, like fear or battle lust, or for moral motives like vengeance and anger. Battlefield robots will always act ‘morally’, that is in conformity to the rules of behaviour they have been programmed to follow.[10]

Overlooking for now the extremely narrow conception of morality involved here, the fact is that these machines are neither autonomous, nor moral and further that this is precisely the point. In consequence of this lack of autonomy, the large-scale use of battlefield robots would concentrate in the hands of a few persons, army commanders and programmers, the power to decide and the power to act morally or not. For Arkins' robots are planned to have a feature, under the responsibility of officers, that allows the robots to bypass their 'ethical limitations', that is, to act in contradiction with the laws of war or rules of engagement, whenever those officers may judge it to be necessary. Ultimately the moral decision lies with humans, commanders or field officers. However, and in spite of the structure of the army's chain of command which in principle should always make it possible to determine responsibility in cases of atrocities or war crimes committed by 'ethical battlefield robots', it is likely that such incidents will be treated as accidents or malfunctions. No one will be responsible.

Similarly, self-driven vehicles endowed with 'ethical' rules will neither be moral nor autonomous but will remain ordinary machines that abide by the rules they have been programmed to follow. The Moral Machine Experiment argues that the rules governing their behaviour in situations where they may cause harm should be those that facilitate the social acceptance of this new dangerous technology, while

[10] Note in passing that the great atrocities of the 20th century were mostly not perpetrated under the influence of strong passions, but simply "following orders". See for example, Z. Bauman. (1989). *Modernity and the Holocaust*, Cornell University Press. ; C. Browning. (2001), *Ordinary Men : Reserve Police Battalion 101 and the Final Solution in Poland*, London : Penguin Books. ; P. Dumouchel. (2015)., *The Barren Sacrifice An Essay on Political Violence*, Michigan State University Press.

avoiding the fundamental issue of responsibility. There are at least two difficulties here. First, it is not clear that the measures, or rules that facilitate the social acceptance of a dangerous new technology are those which afford the public better protection. Especially when they are determined on the basis of unrealistic scenarios of rare events, it is unlikely that they can in anyway address the issue of public security.

The second difficulty is that self-driven vehicles programmed to comply with our 'global ethical preferences' as 'discovered' by this experiment or to comply with any other 'ethical rules' do not have any moral autonomy and therefore do not and cannot have any moral responsibility. However, referring to them as ethical machines and arguing that they should be treated as such fosters the myth that they can be and are responsible. What is involved here is not only an egregious philosophical mistake. As is demonstrated in the case of battlefield robots, whether or not we have the technical ability to create artificial agents that have moral autonomy, the fact is that we do not want to make such machines. We want robots to unfailingly apply the rules, moral or otherwise, that we program them to follow. This desire is not limited to the case of battlefield robots. As the authors of the 'Moral Machine Experiment' write, "we can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; *and to make sure that machines, unlike humans, unerringly follow these moral preferences*" (2018 : 63; italics added). A morally autonomous agent, however is one who can do otherwise, who can act immorally, and it is precisely that freedom or leeway that makes the agent responsible for his action.

Some may ask : "would it be better then for autonomous machines to act immorally?" It is however not necessary to answer such ill-conceived loaded question[11] to show what is involved here.

Attributing moral responsibility to non morally autonomous artificial systems is equivalent to attributing it to no one. In the case of self-driving vehicles it means that no one would be responsible should an accident, fatal or otherwise, happen because of a choice taken by the machine. There are good reasons to want to avoid such a situation of perfect irresponsibility.

One is legal. Currently when a person, individual or enterprise, chooses to adopt a dangerous technology and an accident happens, even if all safety rules and measures have been diligently followed and safety inspections regularly made, that person is responsible. The extent of his or her responsibility will vary depending on many circumstances, but it is clear where, at first sight, the responsibility lies : with the agent who decided to adopt this dangerous technology. It is hard to see why it should not be the same for self-driven vehicles? Why should an autonomous car or truck partake in the legal responsibility of an accident it caused any more than an exploding electrical transformer does? Notwithstanding that 'ethical rules' may have been implemented in the autonomous vehicle, it does not have any more responsibility in what happened than the transformer does.

Another, closely related, reason is social. If operators and owners of self-driving vehicles are held responsible in the case of accidents caused by their machines, they can be expected to exert pressure on manufacturers to produce safer machines, to correct faulty software, and to invent better models. Attributing responsibility to the machine or questioning the moral value of the ethical rules that have been

[11] Loaded in the sense in which the question, "when did you stop beating your wife?" is loaded, it implies that you did. Ill-conceived because what it essentially asks is : "do we want to build morally autonomous artificial agents?" Something which is quite different from how we should consider the non morally autonomous artificial agents which we currently build.

implemented in them will weaken this social feedback mechanism. In the first case, because by letting everybody off the hook it reduces the incentives of manufacturers and designers to invest in making safer models. Especially if the existing ones tend to sacrifice those we have decided 'should die' : the elderly and those who prefer to walk or drive alone. Questioning those 'moral preferences' is unlikely to help either because it displaces the space of discussion from the search for better technology to 'what are we ready to accept?' In consequence, moral and ethical machines as presently understood threaten to become a black hole out of which social responsibility can never escape once it has fallen into it.

There finally are ethical reasons. Alexei Grinbaum recently proposed an interesting and original solution to the question of how autonomous vehicles should respond when faced with the type dilemmas encountered by players of the "Moral Machine Experiment". Self-driven cars, he argues, are simply machines, attributing to them moral responsibility and the power of making moral decisions is a category mistake, and from a technological point of view a misrepresentation. What should these machines do then when such situations arise? The choice should be random. This solution is, technically, perfectly adequate and feasible and it has the advantage of not attributing to the machine an ability which it does not have.(A. Grinbaum. 2018. pp. 103 -156.) Some may object, to the opposite that this would be perfectly irresponsible, that we cannot allow randomness, luck or chance to decide issues of life or death.[12)]

In fact, many reasons conspire to make this an excellent solution.

[12)] They forget, or perhaps do not know, that, original as it may seem in the context of modern cutting-edge technology, randomness is a very ancient means of resolving unusual issues of life or death.

First, it is appropriate to the real capacity of the machines in question. Instead of attributing to self-driven vehicles an ability which they do not have, it recognizes the limit of the complex autonomous system which we presently create. Second, we should remember that such situations are rare and that in philosophical ethics the trolley and other similar dilemmas are not seen as bearers of moral knowledge, but as situations which reveal the limits of our moral intuition. They are not deployed to teach us the right solution, but to show where our moral knowledge ends, and our intuition fails. When the dilemmas are viewed in this way, randomness as a solution is pretty much where we (humans) stand. We do not know and can find no clear rule. Furthermore, it could be argued that this is precisely what the results of the 'Moral Machine Experiment' reveal since for six out of 9 situations no clear preference emerges, precisely as if the choices were random… (2018 : 63) A third, and closely related, reason is that adopting randomness as a solution acknowledges these limits rather than imagining that technology can allow us to transcend them. The moral problem is ours and we should not imagine that machines can solve it for us, that they will allow us to escape it. If one days — and why nots — we do succeed in creating truly moral machines, they will not be a solution that allows us to escape from the limitations of human morality. Migrants, coming from much further than any human foreigner, true moral machines will create a host of new ethical problems and difficulties, but they will also be for us the occasion of moral growth and of new ethical knowledge.

References

Arkins, R. (2009). *Governing Lethal Behavior in Autonomous Robots.* Boca Raton, Flo : CRC Press

Awad, E. Dsouza, S. Kim, R. Schulz, J. Henrich, J. Shariff, A. Bonnefon, J-F. & L. Rahwan, (2018, November). The Moral Machine Experiment, *Nature 563*, p. 59-78. https://doi.org/10.1038/s41586-018-0637-6

Bauman, Z. (1989). *Modernity and the Holocaust.* Ithaca : Cornell University Press.

Browning, C. (2001) *Ordinary Men : Reserve Police Battalion 101 and the Final Solution in Poland*, London : Penguin Books.

Dumouchel, P. (2015) *The Barren Sacrifice An Essay on Political Violence*, East Lansing : Michigan State University Press.

Foot, Ph. (1967). The Problem of Abortion and the Doctrine of Double Effect, *Oxford Review 5*. p. 5-15

Grinbaum, A. (2018). *Les robots et le mal.* Paris : Desclée de Brouwer.

Guillebeault, S. (2019) *Le Bon, La Brute et le Truand. Ou comment l'intelligence artificielle transforme nos vie.* Montréal : Druide

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law*, London : Edward Elgar.

Scheutz, M. (2017. 12. 28). The case for explicit ethical agents, *Ai Magazine 38*(4), p. 57-64 DOI : 10.1609/aimag.v38i4.2746

Suchman, L (2007). *Human-Machine Reconfiguration*, Cambridge : Cambridge University Press.

Received : 31 Aug. 2019, Evaluated : 15 Sept. 2019~30 Sept. 2019, Accepted : 18 Oct. 2019