The automation of ethics: the case of self-driving cars

Raffaele Rodogno, Marco Nørskov
Forthcoming in C. Hasse and D. M. Søndergaard (Eds.) *Designing Robots – Designing Humans*. Routledge

Abstract
This paper explores the disruptive potential of artificial moral decision-makers on our moral practices by investigating the concrete case of self-driving cars. Our argument unfolds in two movements, the first purely philosophical, the second bringing self-driving cars into the picture. More in particular, in the first movement, we bring to the fore three features of moral life, to wit, (i) the limits of the cognitive and motivational capacities of moral agents; (ii) the limited extent to which moral norms regulate human interaction; and (iii) the inescapable presence of tragic choices and moral dilemmas as part of moral life. Our ultimate aim, however, is not to provide a mere description of moral life in terms of these features but to show how a number of central moral practices can be envisaged as a response to, or an expression of these features. With this understanding of moral practices in hand, we broach the second movement. Here we expand our study to the transformative potential that self-driving cars would have on our moral practices . We locate two cases of potential disruption. First, the advent of self-driving cars would force us to treat as unproblematic the normative regulation of interactions that are inescapably tragic. More concretely, we will need to programme these cars' algorithms as if there existed uncontroversial answers to the moral dilemmas that they might well face once on the road. Second, the introduction of this technology would contribute to the dissipation of accountability and lead to the partial truncation of our moral practices. People will be harmed and suffer in road related accidents, but it will be harder to find anyone to take responsibility for what happened—i.e. do what we normally do when we need to repair human relations after harm, loss, or suffering has occurred.

1. Introduction

The idea that robots or machines be considered as moral agents has recently awoken interest (Gunkel 2012, Coeckelbergh 2014). There is still widespread skepticism about the possibility that they can become morally responsible agents, worthy of praise and blame for their actions, on a par with healthy human adults (Rodogno 2016a). Even if robots did lack moral agency, however, they may still have the capacity to perform actions that affect humans and their well-being, that is, actions that occupy a central place in the realm of ethical evaluation. What is more, these machines' doings will look much more like actions than like happenings such as (unprovoked) earthquakes or avalanches. Unlike the latter, the machine's doings will in some cases be the result of a process in which an artificial-agent represents to itself, say, loss of human life as the likely outcome of a certain course of action and then actively selects that course of action. Irrespective of their capacity for moral agency, some of these machines will face situations in which they will have to make a moral call, or at least, what we humans would understand as such if we were the ones who had to make the call or who were interpreting it from our everyday standpoint.

The field of machine ethics, then, has a rationale that is independent of deeper philosophical questions about the very possibility of moral agency for machines. Within this field, discussion has been conducted at two distinct though connected levels. Most generally, researchers have been asking whether the introduction of artificial agency in certain social realms (elderly care, warfare, medical care, children education, etc.) has clear ethical repercussions (Nourbakhsh 2013; Rodogno 2016b; Sharkey & Sharkey 2012 and 2013; Sharkey 2016; Sparrow 2002, 2007; Sparrow and Sparrow 2006; Sullins 2012). Here the question is whether it is desirable and morally permissible to introduce in society technologies such as self-driving vehicles, warbots, or machines that deliver medical decisions and services. At another level, others have asked what substantive ethical principles should be programmed into these machines (Bringsjord & Sen 2016; Govindarajulu & Bringsjord 2015; Wallach and Allen 2009). Should they be allowed to reason morally in consequentialist or non-consequentialist terms.[1]

[1] On a standard understanding, consequentialism claims that the moral rightness of an action is entirely determined by its consequences on the well-being or utility of the individuals that it affects. On this view, an action is right if, and only if, it has the best consequences understood in this sense. At its most basic, non-consequentialism is

Obviously, these two types of questions are connected. If, for example, we decided to programme fairly simple-minded utility maximising principles into our self-driving vehicles or warbots, we can expect that their operation will have a different impact on society than if a whole different set of normative principles had been chosen instead. And, of course, we can expect our evaluation of the impact of this technology on society to vary accordingly. Hence, depending on how utility gets spelled out, a pure consequentialist algorithm might always select the option that minimizes loss of life, irrespective of whose lives they are. It would not matter whether the car would savevpassengers as opposed to pedestrians; children as opposed to adults; those responsible for bringing about the situation (e.g., by jaywalking) as opposed to those who are not; and so on. This will be sufficient ground for some to be extremely critical of this technology or, at least, of this particular consequentialist version of it.

The discussion on which we are about to embark situates itself somewhat in between these two levels. While ultimately attempting to describe and evaluate the impact that the introduction of a specific new automatizing technology would have on our moral life, it will do so not by focusing on the specific substantive decision principles that machines should employ in ethically charged situations, but by focusing instead on the very fact that a decision with ethical implications would be taken by a machine and presented to us as being a legitimate decision.

In what follows, we concentrate in particular on the case of self-driving cars, also known as driverless, autonomous, or robot cars. This choice is dictated in part by the fact that the potential problems introduced by this technology come with a certain degree of urgency, given the importance of cars to our mode of life, and the fact that this technology, though not yet fully operational, is already in our midst. As we discuss this particular case, we should however keep in mind that that the lessons drawn from it might ultimately be extended to other cases, a point to which we return in our conclusion.

understood as denying that only consequences matter morally (the actions themselves independently of their consequences may matter, for example) and that, therefore, the right action is not necessarily the action with the best consequences.

Our argument is in two parts. We start off in Section 2, by describing a set of moral practices, which we hope many readers will find familiar. We focus in particular on a rather neglected aspect of these moral practices, namely, the idea that they are responses to/manifestations of specific features of human moral agency and its context. This section is articulated independently of its application to the ethics of self-driving cars. Yet its articulation is crucial to the argument that follows in the second part of the argument. Our contention there is that the ethics of self-driving cars need to be articulated against the background of moral practice as we describe it. In this part, then, having introduced the crucial technological aspects of self-driving cars (Section 3), we identify two cases of potential disruption (Section 4). First, we show how the advent of self-driving cars will force us to deal with moral dilemmas as if they enjoyed clear and agreed moral solutions. Moral dilemmas, however, are tragic and have no easily agreed solutions. Ordinary moral practice has ways of dealing with them that respect their tragic nature while allowing the occurrence of the restoration of disrupted relations. The advent of self-driving cars will obstruct the functioning of this particular aspect of moral practice, or so shall we argue. Second, this technology will contribute to the dissipation of accountability. While people will still be harmed and suffer as a result of road related accidents, unlike what happens today, it will be harder to find anyone to take responsibility for what happened and do what we normally do when we need to repair human relations as the result of harm, loss, or suffering.

2. Three features of moral life

Moral practices regulate social behavior by way of norms that, among others, specify what we ought to do and not to do, including when someone fails to comply with these norms. Hence, for example, we ought not to harm the innocent and when someone does, it becomes permissible for others to act towards the violator in ways which would be impermissible had the violation not occurred (express blame, exact punishment, etc.). This picture may suggest that moral practices are fairly straightforward, regulated by clear rules with uncontroversial applications. As any mature moral agent would know, however, it would thoroughly unrealistic to characterize moral

life in these terms. As common experience reminds us, figuring out what we ought to do is often a complicated affair, and when we do figure it out, we may well fail to be motivated to do the right thing. What is more, moral action involves reacting to others. Yet others' minds are not immediately open to scrutiny: we cannot properly react to their undertakings until we understand their reasons. Things are more complicated yet. The norms that are supposed to guide and regulate behaviour do not enjoy unanimous agreement, even within smaller groups. Furthermore, even if largely agreed upon, these norms may fail to regulate behaviour exhaustively because life presents us with endlessly varied situations and at times with situations that defy univocal regulation. In what follows, we describe these (realistic) features of our lives under three different headings. More importantly, we would like to suggest that these features are so integral to our experience that they do in fact shape moral practice itself, which can then be envisaged as a way of dealing with them, or as their expression or manifestation.

2.1. Limited motivation, cognitive powers and physical abilities

In our capacity as agents, we are beings that make errors, including moral errors, due to our limited cognitive, motivational, and physical capacities. Sometimes, for example, we simply fail to master the motivation to do that which we ourselves see as the right thing to do. Other times, the motivation might well be there but we fail to figure out what is right perhaps due to lack of time or cognitive resources. Furthermore, we may well know what is right and be motivated to do it, while being unable to do so due to our physical limitations—e.g., being too slow to react. Note that these limitations define us not only as agents committing wrongdoing, but also as agents responding to potential wrongdoing. We might not always know whether a perceived wrong was really intended by its agent, whether she was fully aware of the circumstances that led to her action, whether she should have known better, whether someone else or something else was constraining her freedom of action, and so on. Responses to perceived wrongdoing such as blaming, applying various forms of sanction, exculpating, may well be misplaced and have therefore to be mediated by a process aimed at understanding the agent in the circumstances of her action.

This process of elucidation leads to the assignment of moral responsibility or accountability by probing the extent to which the agent was justified in doing what she did. This process is a central pillar of moral practice. We seek justifications from others when we perceive them as having acted wrongly; and we offer justifications to others (and ourselves) when we feel that we have done something wrong or when others think that. Given our limited capacities, we cannot simply assume that a behaviour that appears to be in violation of a moral norm is worthy of blame and sanction. Even what at first looks like homicide could in fact be justifiable as an instance of self-defense. Actions are made intelligible as the kind of actions that they are (an instance of intentional homicide, manslaughter, self-defense, an act of vengeance, etc.) by being placed in a narrative, by being seen as part of the history to which they belong. Actions typically fail to carry their intelligibility on their sleeves for the immediate grasp of others. Creatures like us need time to understand them. That is why the initial reaction that accompanies a perceived wrongdoing or violation involves a demand for justification: "why did you do that?". The answer to this question is often arrived at by answering other questions: Did the agent really do what she appeared to be doing? Did she really mean to do this? Why? Why to me (or this other person)? What is the nature of relation and history between her and me (or this other person)? Was she fully aware of the consequences? Was she being coerced in some way or other?

To illustrate suppose that a pedestrian was hit by a car as she walked on the pavement, and, as a result, broke her leg and shoulder. Why did the driver do that? Was his action intentional or unintentional? If intentional, was there an intention to harm the pedestrian deliberately? If so, why? Do the pedestrian and the driver know each other? What is the exact nature of their relation? Was it the desperate move of a spouse trying to stop her beloved from enrolling in the army or from leaving him for someone else? Or was the intentional action not deliberately harmful but the only option available to the driver not to harm someone else, say, a child who suddenly jumped into the street while running after her ball? Even then, however, perhaps the accident might have been avoided had the car been driving within the speed limit. If unintentional, was the action due to (culpable) negligence, e.g., failing to have had the car breaks checked? Or was it due to being distracted as a result of texting on the mobile phone or of driving under the influence?

At this point, the agent—in the case above the driver—offers her reasons, which may or may not persuade others. She will offer her perspective on the event: given what she knew or failed to know and the features peculiar to her situation, she might try to show that her action was a reasonable action to take or perhaps not as unreasonable as it might at first appear. As agents who are asked "Why?", we explain ourselves to others in the hope that they come around to appreciating and sharing the force of the reasons that motivated us. Sometimes our explanation may be so successful as to exculpate us from any wrongdoing: we are not the proper object of indignation or blame after all. Other times, our reasons will at most mitigate our fault. We are to blame but not as much as it appeared at first.

Yet other times, agents do not have any good reasons to offer for their behavior, and that even by their own lights ("I simply don't know why I did that"). After the fact, some agents can even see the wrongness of their action, its unacceptability from the point of view of others who had to suffer the action or its consequences. But even in these cases, moral practice offers wrongdoers a remedy in the form of apologies, other expressions of remorse and shame, as well as action tendencies aimed at repairing the wrong and/or restoring relations. Similarly, those wronged will feel indignation or resentment and will generally be entitled to express blame and exact retribution or reparation, or have them exacted for them.

If we had the capacity always to know and do what is right, there would be no need for justification or for processes of atonement. Just as we have limitations as moral agents who initiate action, we have limitations as agents who respond to the actions of others. In both cases, the practice of justification comes to our rescue. Its process of intersubjective sense-making, the assignment of moral responsibility and the practice of reconciliation that it enables, are the expression of these limitations.

2.2. Disagreement and Incompleteness

A second pervading feature of moral life is its lack of explicit and exhaustive guidance. Our moral practices do not contain norms for all the moral calls which we may be asked to make and this for

two kinds of reasons. The norms that regulate moral behavior fail to cover all the possible situations which we may face. Moral life is varied, intricate, and complex. In fact, the norms may often seem to come into conflict with each other, making it difficult to know how to apply them. What is more, even when we thought we knew what to do in light of a norm, there may be disagreement within the community in which we operate, as to the legitimacy of this norm (as applied to this context).

Moral life is complex even in its everyday guise. You may for example wonder whether it is time for you to ensure that your older parents be moved in a home; or whether you should warn your friend that, in your view, her current partner is really not a wise choice; or whether you should keep the extra money that have been wrongly credited to your bank account; and so on and so forth. More to the point, our moral norms may fail to determine what to do when certain (unfortunate) circumstances arise on the road. They may, for example, fail to tell us how to balance the conflicting rights or interests of various people involved, as when our only option is either hitting the child who suddenly runs into the street or hitting the pedestrian who walks on the pavement.

We do not have specific and precise moral norms for all these situations. Even if such norms could be devised, we may well expect members of a community to disagree about them and their application. Taken as a body, the positive moral norms of a community typically fail to enjoy unanimous endorsement and are the object of continuous challenge and negotiation. Challenge and negotiation of norms are more likely to be expected in those areas where regulation pitches the interests or rights of certain groups of individuals against those of other groups (rich vs poor; different ethnic groups; gender) or where regulation has to strike the balance between the most fundamental values of a group (e.g., equality vs freedom). As we will see in Section 4, the advent of self-driving cars is likely to usher controversial regulation, likely to be the object of fundamental disagreement.

Moral practice, however, has some ways of coping with the normative indeterminacy and disagreement discussed here. The justificatory exchanges described above are also obviously relevant here. When asked why she did what she did, an agent may be able to show how from her

perspective it was not at all clear what norm, if any, applied to the situation. Other times, this process will allow to see that failure to comply with a norm does not stem from bad will or weak will but from a fundamental disagreement with the relevant norm (think about conscientious objection). Our reaction to the latter is different to our reaction to the former. This is not to say that we will always be able to reach agreement with our moral interlocutors, to see how the agent's action was a reasonable action from her point of view. In these cases, societal practices typically delegate the task of adjudicating disputed cases to individuals that are endowed with the relevant authority.

2.3. Lack of order

According to some belief systems the flagrant injustices and imbalances that we witness or experience in this world are eventually redressed by a divinity or by other forces, at a later point in time (think about the notions of reincarnation or karma) or in another dimension (think about the Christian notions of hell and paradise). In this way, the universe is, as a whole, in balance, and the embodiment of a perfect moral order. Lack of moral order, however, is what we witness and experience in our lives. We typically do so from our perspective as *recipients* of injustices caused by other moral agents and non-moral agents (e.g. natural catastrophes). What is of interest here, however, is that lack of order can also be experienced from our perspective as *agents*. Even when we think and act impeccably, life may well place us before tragic or dilemmatic choices, in which all the options available to us as agents are bad options. When in such predicaments, even the best course of action will not be good enough; someone will suffer or be wronged as a result of our agency. In this sense, lack of moral order involves "innocent" tragic agency –-innocent because agents find themselves in such situations through no fault of their own, whether the situation is itself the making of other human agents, or an unfortunate outcome for which no human agent can clearly be blamed (the result of bad luck, or fate, or divine intervention).

Moral dilemmas are the best illustration of a lack of moral order as we may experience it from our perspective as agents. There is a tradition for discussing important philosophical and ethical issues through the device of moral dilemmas. Notable among "philosophical" dilemmas is the case of

Antigone and the choice forced on her between honoring family and religion at the expense of the rules of the *polis*. Or again, Sartre's (2007) WWII case of a young man who has to choose whether to join the *Resistance* and leave his frail mother to fend for herself; and *Sophie's choice* (William Styron's novel (1979)) in which, upon her interment at Auschwitz, Sophie has to choose which one of her two children would die immediately by gassing and which would continue to live, albeit in the camp. In all these examples, the agent's choice will have tragic consequences for him or herself as well as for others. What is more, it looks as if there are no right answers. Can we really blame any of these agents for choosing whatever they end up choosing or, indeed, for choosing the other available option, had they decided for that instead?

In circumstances such as this, it is difficult to know what is right and what is wrong; this, however, is no ground for absolute pessimism. First, the fact that we do not have clear answers now does not exclude that we may through various dialogical processes agree on some answers at a future point. Second, and most importantly, we can once again envisage moral practices as a manifestation of this particular feature of moral life. As an illustration, consider the role of two important moral emotions, namely, remorse and regret.

Remorse (or guilt) is paradigmatically invoked as the emotion that we feel when we experience ourselves as having acted in ways that violate an internalized norm, one that is important to us (Deonna and Teroni 2008). Remorse is often also invoked as characterizing dilemmatic choices such as the one by Sophie or the young man above. Some argue that in such dilemmatic choices, it is not appropriate or fitting for agents to feel remorse, for, after all, they did not really have the choice not to do something bad that violated an important norm. Yet, at the same time, it is perfectly intelligible that agents do experience such situations in these terms. It is just natural to read that Sophie, who escapes from Auschwitz alive and moves to New York, is ridden with guilt, and plunges into depression and alcoholism and eventually takes her life. Similarly, we would find it quite intelligible to learn that the young man from Sartre's story was guilt-ridden for many years upon discovering that his old mother died in solitude shortly after he joined the *Resistance*. Remorseful subjects want to undo their actions, to make it good again. Their action tendencies typically involve the idea of making amends (that is why guilt can be a dangerous emotion in cases such as Sophie's, for it is impossible in her case to undo what has been done; her remorse is like

an open wound, incapable to heal). So, even though the world in its lack of moral order draws us into difficult situations, we have ways of dealing or coping with it. Through our guilt we may sometime be able to expiate, atone, and resume normal existence.

The case of regret is different in so far as it does not involve the idea that, if faced with the same situation, one would not do the same. Imagine an unfortunate railway operator who finds herself in a tragic situation: she notices that the five passengers on a runaway trolley will be killed unless she flips a switch and shifts the trolley onto another set of tracks. Unfortunately, however, a rail worker is currently working on these tracks. If the operator flipped the switch, he will certainly be killed. What is the operator to do? This is the (in)famous "trolley problem", a dilemma used to make a variety of claims in both philosophy (Foot 1967; Thomson 1985; Kamm 2015) and cognitive science (Green 2013; Mikhail 2013). Unlike these discussions, however, we use this dilemma here to illustrate the role of regret in our moral lives, given the lack of moral order that characterizes the reality in which we operate. Whatever your specific moral intuitions about this case, we would argue that most people should find it hard to consider the operator blameworthy either way. We should condone the operator if she did not have it in her to actively flip the switch. This would dramatically change her causal role in bringing about the death of someone: as a result of her intervention, the rail worker would never go home to his family again. However, we should also understand her reasons if she did flip the switch. After all, by doing so, she would save five lives and sacrifice only one.

Either way, however, we would think that the situation is regretful and would expect the operator to feel this much. Even if she believed that she did what was right, and were disposed to do the same thing if the situation arose again, we would think less of her if she failed to feel regret (Williams 1981, 27-30). In fact, we can easily imagine how her lack of regret, if perceived by the friends and relatives of those who died in the accident, may ignite (or fail to assuage) their indignation and sense of justice. Sincere expressions of regret serve precisely to signal to others (and to the agent herself) that she is saddened to have had to act (or react) in a certain way, thus being the vehicle of misery, in a tragic situation, which the agent is not herself blameworthy for bringing about.

As discussed in the last subsection, life presents us with many difficult or hard cases. Not all of them should be seen as dilemmas that have no right answers. The point here is rather that moral practice can be envisaged as a way of dealing with the difficult features of moral life discussed above. It allows us, human agents, and those who suffer our actions, to enter into a dialogue in order to understand each other and the nature of our actions; it allows us to reconcile and atone after relations are interrupted by our moral errors or by the complexity, indeterminacy, and lack of order of moral life. It is against the background of this particular view of moral life that we will assess the ethical impact of self-driving cars in Section 4. First, however, we need to present some key features of this particular kind of technology.

3. Training cars

Part of the ethical justification for deploying self-driving cars is that, by being much less imperfect drivers than we are, they will inflict less damage on human beings (and other creatures) than human beings themselves. By being cognitively and physically more resourceful in reacting to their environment, these machines will have options that are not available to humans placed in the same circumstances. Hence situations that would be dilemmatic for humans may not be so for the machines, for their greater resources would make more and better options available to them. It is generally agreed, however, that despite their greater powers, these artificial agents too will eventually encounter dilemmatic situations.[2] The question then arises as to how we should design them so as to cope with these cases.

[2] In the relevant literature on self-driving cars, the discussion around such dilemmatic situations has so far focused on a particular type of dilemma, namely, the trolley problem introduced above. While most researchers seem to agree that autonomous cars will face this type of dilemma, there is less agreement concerning its significance for the ethics of self-driving cars. A number of authors (Lin 2015, 78; Wallach and Allen 2009, 14; Bonnefon et al 2015, 3; Sütfeld et al. 2017; Goodall 2014a) take these cases to be central to the ethics of autonomous cars. Others dismiss them for different reasons. Nyolm and Smids (2016) rightly claim that the ethics of accident algorithms for self-driving cars and the trolley problem differ from each other in some important respects, among which the time

Before answering this question, it is worth considering how these machines are actually being designed. The self-driving cars being tested today are all equipped with some version of machine learning. What distinguishes machine learning from the earlier forms of AI (symbolic AI) is, in short, the capacity for learning and originality. That is, we move from machines that can do whatever we know how to order them to perform, to machines that can learn on their own how to perform a specific task. As F. Chollet (2017) succinctly puts it:

> In classical programming, the paradigm of symbolic AI, humans would input rules (a program), data to be processed according to these rules, and out would come answers. With machine learning, humans would input data as well as the answers expected from the data, and out would come the rules. These rules could then be applied to new data to produce original answers. A machine learning system is "trained" rather than explicitly programmed. It is presented with many "examples" relevant to a task, and it finds statistical structure in these examples which eventually allows the system to come up with rules for automating the task

For instance, Chollet goes on to write, if you wish to automate the task of tagging your vacation pictures, you could present a machine learning system with many examples of pictures already tagged by humans, and the system would learn statistical rules for associating specific pictures to specific tags. In short, Chollet concludes, in order to do machine learning we need three things: (i) Input data points, e.g. the picture files; (ii) examples of the expected output, e.g., tags such as “beach”, “sea”, and so on; and lastly but most importantly (iii) “a way to measure if the algorithm is doing a good job, to measure the distance between its current output and its expected output. This is used as a feedback signal to adjust the way the algorithm works. This adjustment step is

pressure and conditions under which the morally relevant decision would be taken in each case (a point to which we return below). Similarly, Goodall (2016) seems to think that it is better to include trolley cases under the much larger rubric of risk management. Finally, Bringsjord and Sen (2016, 781) envisage a multi-agent AI future in which these “silly” ethical dilemmas will be very rare because prevented by the very creative and powerful technology and where the real ethical difficulties will be of an entirely different kind. Our stance is that this type of dilemmas has at least some instrumental significance within the ethics of self-driving cars. In particular, it serves the purpose of focussing our attention on the fact that some difficult and potentially controversial ethical decisions will need to be addressed at some point in the design of autonomous cars.

what we call "learning"" (Chollet, 2017) and this function is what we call “loss function” or “objective function”.

Applied to the context of self-driving vehicles this means that:

> Instead of supplying the vehicle with a fixed evaluation scheme from which the right action for each situation can be deduced, the programmers feed the software with many traffic situations and specify the correct action for each situation. The program then searches by itself for the best configuration of internal parameters and internal decision logic which allow it to act correctly in all of these situations. Like with us humans, it then becomes difficult to answer the question why the car exhibits a specific behavior in a new situation: no “explicit rules” have been specified; the decision results from the many traffic situations to which the algorithm had been exposed beforehand. (Hars 2016, 4):

As we shall see, the idea that driverless cars are trained by way of an “objective function” and the idea that they somehow make their own rules are of particular interest when evaluating the potential impact of such cars on our lives.

4. Impact

4.1. Controversial extensions

Some of the scenarios that engineers have to envisage are mission critical—i.e. life and death cases or, in the absence of that, cases in which serious harm is involved. More in particular these are situations in which driverless cars are imagined to have only two options open, whereby swerving to one side involves death/harm to one person, and swerving to the other (or not swerving) involves death/harm to another person, where one of the persons involved may or may not be the one sitting

in the car. Alternatively, the imagined scenarios involve the death/harming of different number of people on different sides.

When "feeding" the software with traffic scenarios such as these and with the relevant instructions, programmers must inevitably take some ethical stances. If in a dilemmatic situation the machine predicts the outcome "swerves left" and thus harms the person on the left, the engineer must then either believe that this is permissible or believe that this is the morally wrong answer and that the machine needs training. Remember the role of the "objective function". If the outcome predicted by the machine is far from what it should ideally be (the loss score is high), the machine needs to be "trained" with the correct answers. This is how it learns. This approach presupposes that there are answers, and, in fact correct answers to the moral questions posed by these situations, which the "objective function" can use to train the algorithm. In light of our discussion in Section 2, however, this is not an assumption we can simply make. As a community we may have not yet articulated norms that regulate this type of situations, or perhaps we are deeply divided over what norms to adopt, or even think that some of these situations are dilemmatic and cannot be regulated univocally.

Up to now we have not faced the type of questions that engineers working with driverless cars must face, not, anyway, in the context of car driving. From the point of view of our moral practices, these questions are, to some extent, unchartered territory. Currently, a (human) driver may find herself in such dilemmatic situations in two different kinds of circumstances, one that engages her moral responsibility and one that doesn't. In the first case, the driver is considered at least partly blameworthy for bringing the dilemmatic situation about. This is so when the driver has violated one or more of the traffic rules (e.g., speed limits, driving under the influence, failing to have the brakes checked, etc.) whose aim is precisely to ensure safety on the road and avoid such situations. Note how in this situation the driver is *not* considered to be blameworthy for running over one person rather than the other; if we confronted her, we would not ask her why, for example, she run over the boy on the left as opposed to the woman on the right. We would rather blame her for breaking the relevant traffic rules and, hence, for bringing about the dilemmatic situation in the first place. Once she finds herself in it, there is no time to deliberate and choose which person to

run over, and hence be held accountable for that choice. If there was time for that, there would be time to stop the car and avoid tragedy altogether.

Under the second kind of circumstances, the driver is not considered to be responsible for bringing about the dilemma. She has not broken any traffic rules but finds herself in the dilemmatic situation through no fault of her own. Here, she is once again *not* deemed responsible for the actual outcome (running over A rather than B) because, once again, the situation is such that there is no time to deliberate and act but, at best, time to instinctively *react*.

Our engineers, however, are facing a different situation. Unlike human drivers, we can expect self-driving cars not to break the relevant rules thereby generating the dilemma in the first place. Yet, unlike humans, when in the dilemma through no fault of their own, they will still have the power to choose which course of action to pursue. The machine will implement whatever course of action its algorithm was trained to choose, and if the situation that it faces is novel, the algorithm will select whatever course action it predicts to be the correct one. Unlike humans, the car must now have an answer to the question "Should A be run over as opposed to B?".

This is unchartered moral territory. What are the answers to this kind of question? Where are they to be found? In a recent study Sütfeld et al. (2017) have studied the “decisions” of human agents facing tragic situations under time pressure, i.e., when having only either 4 seconds or 1 second to decide. Their study is premised on the idea that the decision model on which to base the algorithm of self-driving cars should be based on precisely this type of reactions.[3] But why utilize instinctive, untrained, or unreflective reactions rather than reflective ones? After all, this technology provides us with the unprecedented opportunity to extend our reflective agency to areas in which only instinctive reactions were possible before. The question, however, is whether our moral powers are up to the new task that this technology now forces us to consider. There is reason to be sceptical. After all, the advent of this technology has changed nothing to the dilemmatic and tragic nature of these situations. The fact that algorithms can now be programmed with “correct” answers, does not mean that there now suddenly are correct answers, or that these situations are any less

[3] Note, however, that Sütfeld et al. (2017, p.12) see their algorithmic decision model as an alternative to machine learning rather than as a particular instantiation of it.

dilemmatic and tragic. In fact, as indicated by studies such as Sütfeld et al.'s, the engineers' "correct" answer can only be expected to be morally controversial. Their alleged correctness is only going to mask the tragic nature of the situation. In doing so it will likely obstruct the normal course of moral practice, which, as we saw, would in such circumstances engage the restoration of relations by way of expressions of guilt and regret. This, however, is what we will argue next.

4.2. Unwelcome Truncations

We may nonetheless think that this problem can be successfully tackled. We may envisage engaging society in an open and informed discussion concerning the appropriate training for driverless cars. This informed discussion would then be followed by a democratic decision. Even a transparent, informed, democratic process, however, will not solve all of the problems. As discussed above, even if society as a body agreed on some rules or principles, these principles will fail to be accepted by consensus and will be resisted and challenged by some individuals or groups. On a topic such as this we may in fact expect important divisions to arise. Should we, for example, train the algorithm to discriminate between different (groups of) individuals (the young over the old? the healthy over the ill? etc.). That is bound to be controversial. Faced with these difficult questions, we may as a society decide not to discriminate among individuals in this way and have a random lottery select on each occasion which individuals to harm or not to harm[4]. But this is also bound to be controversial, for many will think that momentous decisions, involving loss of life, cannot be taken by the flip of a coin.

Suppose, however, that as a society, we nonetheless democratically decide to train the relevant algorithms in a specific way and implement our decision. Consider now what would happen whenever an accident takes place and someone is harmed or killed. The victims, those close to them, and those who represent their interests will want to understand whether someone was

[4] As in the case of the allocation of scarce dialysis machines discussed by Johnson (2001, p. 42-43). The machines were at first distributed on the basis of utilitarian principles, which, however, had the effect of favouring certain types of individuals, e.g., doctors, over others, e.g., criminals. This was deemed controversial and, as a consequence, the distributive principle was changed to a randomized one.

responsible for infringing the relevant norms, and whether there were any reasons that would excuse or exculpate such an act. Even if driverless cars will not be responsible for bringing the dilemmatic situation about (they will not exceed the speed limits or drive under the influence), once in the situation, they will actively select one course of action over another, thereby harming one person rather than another. Unlike the instinctive reactions of human beings placed in the same circumstances, a certain degree of premeditation is at play here. The victim and/or those close to her might feel particularly hurt and entitled to some expressions of regret if not remorse, something which, as we saw, plays an important role in quelling the sense of injustice or indignation and in restoring relations.

But who is going to feel and express regret or remorse now? The passengers in the cars might not be inclined to do so. They were not the ones driving the car after all. We imagine that they will rather take themselves to stand to car victims in a relation similar to that in which passengers riding on a train stand with regard to potential train victims. Much of the moral involvement that derives from actually being the driver, the active cause of harm or death, will likely disappear.

We may then turn to car producers. But in the scenario envisaged here, the autonomous car has acted in accordance to the norms democratically approved by the regulators and has not malfunctioned. Producers will in these circumstances feel no more inclined to attune than they do today when a human driver runs over someone while driving one of their cars. Finally, we might want to turn to society itself, which in this scenario is ultimately responsible for determining the cars' behavior. But who, or what official body should express regret and remorse? And why should any such body express regret for something that was after all democratically decided and was hence endowed with legitimacy?

What this points to is a partial truncation of ethical life. As argued in Section 2, assigning moral responsibility by way of probing the agent's justification for her action is a central pillar of moral practice. It is the first step towards the restoration of relations after the occurrence of harm or loss. Now, however, no one is there to take responsibility and no one is likely to express remorse or even regret, and that is not because harm or loss no longer occur. As we are about to see, however, this is not the only truncation likely to affect moral life.

As noted in Section 3, the advantage of machine learning as opposed to classical programming is precisely the fact that it might select courses of action for new situations, ones which have not been met before and, hence, have not actively been trained ahead of releasing the cars on the road. As explained above, programmers input data as well as the answers expected from the data, and out come the rules. These rules are in a mathematical language, not the language of reasons that we use when attempting to justify our actions to each other. This may not be a problem when the situation faced by the car is identical to one that was "trained" with the "correct" answers. Yet, once trained, new situations (data) is processed and categorised automatically without operator intervention. This may lead to the cars selecting courses of actions whose rationale is pretty obscure to us. Some such actions may well be perceived as wrong or harming and all we will have to go by to understand and justify them is a machine-generated algorithm; a far cry from our current moral practices.

Artificial machine reasoning based on the probability theory that animates machine learning, is not always the suitable interface to human reasoning and intuition (cf. Hood 2013, 163)[5]. If machine "*phronesis*" is communicable to us by ethical artificial decision makers but not adequately intelligible to us—e.g. due to the complexity of the underling calculus of probabilities, which may be beyond our intellectual capacities or simply infeasible for us to comprehend within the given time or resources (Mittlestadt et al. 2016, 6)—we face the risk of being deprived of the essential moral practice of understanding our own, and the actions of the other. Our practice of justification will suffer as a result and so will our capacity to restore relations.[6]

4.3. Further considerations about responsibility

We have so far imagined that society would regulate the introduction of this technology centrally, democratically, and univocally. We could, however, imagine that the regulation of self-driving

[5] The results of the Monty Hall Problem are for instance not intuitively accessible to many of us (ibid., 160-163).
[6] Bringsjord and Sen (2016, 773) also claim that self-driving cars must be able to cogently explain and justify in a natural language such as English or German their actions. This task, they go on to argue, requires that the underlying technology be logicist and hence unlike machine learning.

vehicles' ethical settings was left in the hands of producers and consumers. The former could give consumers the option of, say, choosing the settings that saved the passengers at all costs rather than those that selected whom to avoid by the 'flip a coin'[7]. In this scenario, as consumers, we would exercise our ethical judgement and then subscribe to it. If any harm or loss occurred as a result of our choice, it should be easier to hold us responsible and, if harm occurred, we would have someone to engage with in order to seek justification, compensation, and the restoration of relations.

Note, however, how, the exercise of choosing what to do may be a rather detached and abstract one, which we may carry out while drinking a cup of coffee at the car dealers' shop. This context may well fail to bring forth the immediacy of the situation with its moral details. Even if we were well aware of the relevant moral details, however, as argued above, we would still lack clear answers. So far, the acquisition of a driving license does not involve any specific ethical training. Decisions on, for instance, whether or not we should run over the little girl in order to save our passengers, or other trolley case-like examples, are not something we are even suspected to have made up our minds about before being allowed to sit behind the wheel. In the scenario imagined here, however, we may find ourselves forced into taking a rational and premeditated stance on ethical dilemmas that have no clear and uncontroversial answers. Finally, the choice architecture deprives us of the sort of leniency that would normally be invoked in attempting to justify our action in cases such as this. Suppose you hit a pedestrian but, in doing so, save yourself and the passengers in the car. As you justify your action, under normal circumstances, you would appeal to the fact that your action was not premeditated and determined by involuntary factors such as the physical limits that constrain our reaction times. If you deliberately chose settings that always prioritize the interests of the passengers, however, the outcome would in some sense be the direct result of a deliberate choice you have made. This, we think, would amount once again to placing a strain on moral life as we know it. Do we really wish to make ourselves fully responsible for this type of situations?

[7] On a more general level Rapoport (2016) argues that this freedom of choice may not be so obvious if stakeholders such as insurance companies become involved.

5. Conclusion

In the last section, we have explored the impact on moral life and practice of automated decision making as this technology is incorporated by self-driving cars. Unlike much of the current debate, we have done so while staying away from normative ethical questions about the principles, be they consequentialist or other, which should be used in training these algorithms. We have rather described moral practice as the manifestation of certain features of moral life and used the picture that resulted as the background for the proper assessment of this type of technology.

We have showed how self-driving vehicles will be involved in unavoidable accidents, with no clear-cut ethical 'right thing to do', while yet being designed as if their "actions" were morally uncontroversial. Those in the car, including those we would have formerly identified as drivers, will be inclined to dispense with emotions such as remorse, and regret, as their status becomes similar to that of passengers on a train. Furthermore, the victim/s and people affected by the situation may find themselves in a depersonalized vacuum, where intelligible explanations and justifications for the tragic happening cannot be produced, and where there is literally no human other to blame or to offer expressions of sorrow and regret, which are so important for coming to terms with the tragedy and to restore relations.[8] We argued further that even in the cases in which there might be someone to blame, the introduction of this technology might constrain into choices that we might not want to have.

Our exposition has been mostly descriptive in nature, pointing to the potential disruption of central aspects of moral practice due to automatization. We have in other words remained silent with regard to what should follow, normatively. If suffering and/or loss of life is higher due to the imperfections of human drivers, would it not be better to introduce this technology despite the aforementioned disruptions?

[8] Somewhat similarly to the point made here Coeckelbergh (2016, 754) writes: "Phenomenologically speaking, the user stops being a driver and becomes a passenger, and passengers are not responsible for what the driver does. The driver is responsible, but the driver has been replaced by a machine."

If self-driving cars were the only algorithmic decision-maker in our lives, this may actually be a small price to pay in exchange for the lives and suffering that would be spared. There would, however, be greater cause for concern if we outsourced as much decision-making to machine-learning algorithms as possible—e.g. have algorithms steer our cars, trade our stocks, pick the right person to fire or to give a bank loan to, deliver medical diagnoses, and whatever else is conceivable and technically feasible. Though each of these cases would need individual attention, we suspect that they would have similar effects as the ones described above. The question would then be to understand their cumulated impact on our lives. Our concern is that, while the advent of this technology will not remove the complex features that complicate our moral reality, it will affect our capacity to deal with these features by removing opportunities to justify ourselves to each other, and to reconcile and atone.

References

Bello, P., Bringsjord, S. 2013. On How to Build a Moral Machine. *Topoi* 32: 251–266

Bonnefon J-F, Shariff A, Rahwan I (2015) Autonomous vehicles need experimental ethics: are we ready for utilitarian cars? arXiv:1510.03346 [cs]. Retrieved from http://arxiv.org/abs/1510.03346

Bringsjord, S. & Sen, A. (2016) "On Creative Self-Driving Cars: Hire the Computational Logicians, Fast" *Applied Artificial Intelligence* 30.8: 758-786

Chollet, Francois. 2017. *Deep Learning with Python*. ISBN 9781617294433 https://www.manning.com/books/deep-learning-with-python#downloads

Coeckelbergh, M. 2014. The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, 27 (1), 61-77

Coeckelbergh, M. 2016. Responsibility and the Phenomenology of Using Self-Driving Cars. *Applied Artificial Intelligence*, 30:8, pp.748-757

Deonna, J., Teroni, F. 2008. Differentiating Shame from Guilt. *Consciousness and Cognition* 17 (4):1063-1400

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

Goodall, N. 2014a. Ethical decision making during automated vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board 2424:58–65

Goodall NJ. 2014b. Machine ethics and automated vehicles. In: Meyer G, Beiker S (eds) Road vehicle automation. Springer, Dordrecht, pp. 93–102

Goodall, N.J., (2016) Away from Trolley Problems and Toward Risk Management. *Applied Artificial Intelligence*, 30:8, 810-821,

Govindarajulu, N. S. & Bringsjord, S. (2015), Ethical Regulation of Robots Must be Embedded in Their Operating Systems, in R. Trappl, ed., 'A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations', Springer, Basel, Switzerland, pp. 85–100

Greene J (2013) Moral tribes: emotion, reason, and the gap between us and them. Penguin, New York

Gunkel, D. J. (2012). The Machine Question: Critical Perspectives on AI, Robots, and Ethics. Cambridge, MA, The MIT Press.

Hars, A. 2016. Top misconceptions of autonomous cars and self-driving vehicles. *Thinking outside the box: Inventivio Innovation Briefs* Issue 2016-09. Retrieved on 9 September 2017. http://www.inventivio.com/innovationbriefs/2016-09/Top-misconceptions-of-self-driving-cars.pdf

Hood, B. 2012. *The Self Illusion: How the Social Brain Creates Identity*. Oxford: Oxford University Press.

Johnson, D. G. (2001). Philosophical Ethics. Computer ethics. Upper Saddle River, Prentice Hall**:** 26-53.

Kamm F (2015). The trolley mysteries. Oxford: Oxford University Press

Lin, P. 2015. Why ethics matters for autonomous cars. In Autonomes fahren: Technische, rechtiche und gesellschaftiche aspekte, eds. M. Maurer, C. Gerdes, B. Lenz, and H. Winner, 69–85. Berlin, Germany: Springer.

Mikhail J (2013). Elements of moral cognition. Cambridge: Cambridge University Press

Mittelstadt, B.T., Allo, P., Taddeo, M., Wachter, S., Floridi, L., 2016. The ethics of algorithms: Mapping the debate. *Big Data and Society* 3.2, pp.1-21

Nourbakhsh, I. R. (2013). "Dehumanizing Robots." *Robot Futures*. Cambridge, The MIT Press: 49-63

Nyholm & Smids (2016) The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice* 19:1275–1289

Rapaport, M. 2016. Persuasive robotic technologies and the freedom of choice and action. In: Social Robots: Boundaries, Potential, Challenges. ed. / Marco Nørskov. UK : Ashgate,

Rodogno, R. 2016a. Robots and the Limits of Morality. In: Social Robots: Boundaries, Potential, Challenges. ed. / Marco Nørskov. UK : Ashgate, p. 39-55

Rodogno, R. 2016b. Social Robots, Fictions, and Sentimentality. *Ethics and Information Technology*, Vol. 18.4, 257–268.

Sartre, J-P. 2007. *Existentialism is a Humanism*. New Haven, CT: Yale University Press.

Sharkey, A.J.C. 2016. Should we welcome robot teachers? *Ethics and Information Technology* 18.4, pp 283–297

Sharkey, A.J.C. and Sharkey, N.E. (2012) Granny and the robots: Ethical issues in robot care for the elderly, Ethics and Information Technology, 14, 27-40

Sharkey, N.E. (2012) Automating Warfare: lessons learned from the drones, Journal of Law, Information & Science, 21(2),

Sharkey, N.E. and Sharkey, A.J.C (2013) Robot Surgery on the cutting edge of ethics, IEEE Computer, January, 56-64

Sparrow, R. 2002. The march of the robot dogs. *Ethics and Information Technology* 4: 305–318, 2002.

Sparrow, R. 2007. Killer Robots. *Journal of Applied Philosophy* 24(1): 62-77.

Sparrow, R. and L. Sparrow (2006). "In the Hands of Machines? The Future of Aged Care." *Minds and Machines* 16(2): 141-161.

Styron, W. 1979. *Sophie's choice*. USA: Random House:

Sullins, J. 2012. Robots, Love, and Sex: The Ethics of Building a Love Machine. *IEEE Transactions on Affective Computing* 3(4):398-409 · October 2012

Sütfeld LR, Gast R, König P and Pipa G (2017) Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure. *Frontiers of Behavioral Neuroscience* 11: 122.

Thomson JJ (1985) The trolley problem. Yale Law J 94(5):1395–1515

Williams, B. 1981: *Moral Luck*. Cambridge: Cambridge University Press

Wallach, W. and Allen, C. 2009. *Moral Machines: Teaching Robots Right From Wrong.* Oxford: Oxford University Press.