Essays

Break the chains: a new way to consider machine's moral problems

Paolo Sommaggio, Samuela Marchiori*

BREAK THE CHAINS: A NEW WAY TO CONSIDER MACHINE'S MORAL PROBLEMS

ABSTRACT: 'Moral machines' are entering our society and with them is coming the need for new and suitable moral, ethical and legal regulations. In the paper we analyse Philippa Foot's trolley problem experiment as an example of a situation in which self-driving vehicles will have to make life or death decisions autonomously. In doing so, we investigate the basis for the construction of moral questions both for humans and machines by considering the problem from a philosophical, social and neuroscientific perspective. Following Rittel and Webber's footsteps, we also highlight the fallacies of the deontological and utilitarian traditional 'one-right-answer' approach, where a solution is undoubtedly right or wrong, and claim that moral problems are not, due to their intrinsic dilemmatic nature, resolvable. We then present a different approach on the matter, arguing for the central and creative role of the tragic as a new tool for enhancing autonomous vehicles' approach to moral problems.

KEYWORDS: Trolley problem; self-driving cars; moral dilemma; cognitive neuroscience; experimental ethics

1. Introduction

'Moral machines' are among us[1].

Self-driving vehicles, for example, are entering our society on the promise of a reduction in the number of road accidents, and with them is coming the need for new and suitable moral, ethical and legal regulations[2]. What constitutes the main source of concern is the possibility that such vehicles will find themselves in situations in which they will have to make decisions autonomously[3].

* *Paolo Sommaggio: Associate Professor, Faculty of Law, University of Trento. Mail: paolo.sommaggio@unitn.it. Samuela Marchiori: undergraduate student, Faculty of Law University of Trento. Mail: samuela.marchiori@studenti.unitn.it.* *The article was subject to a double-blind peer review process.*

[1] See W. WALLACH, C. ALLEN, *Moral machines: Teaching robots right from wrong*, Oxford, 2010.

[2] In this paper, we do not investigate the legal implications concerning autonomous vehicles' potential civil or criminal liability, nor do we avail ourselves of sectoral legislation on the matter, as these topics do not lie within the scope of this research. M. WINDSOR, *Will your self-driving car be programmed to kill you if it means saving*

Decisions of this kind are machine decisions, but are they based upon the moral principles of machines or those of humans?

In this paper we do not investigate whether machines may have specific moral principles, but instead we presume that machines will have the same moral principles as humans[4]. 1 Therefore, before investigating the 'moral decisions' of machines, we have to clarify the human approach to this scenario and the rules coming from a human moral system[5].

This is the focus of the paper: what is the best way to construct moral questions for humans and for machines? We will suggest a new way to consider these questions: they are not resolvable problems, where a solution is undoubtedly right or wrong, but we have to assume that a moral problem has no resolution because it is not a question of mathematics but has the same structure as a tragic choice. We think this kind of choice may be the best way to consider the structure of a moral problem for humans and machines in a situation where every choice has a 'cost' in moral terms. In doing this, we will first show that the mimetic moral setting for machines is a human setting. Secondly, we will focus on how the human moral setting is traditionally discussed, using a famous example: the trolley problem. We will then analyse both the utilitarian and the deontological solutions. In the next step

---

more strangers?, in *Science Daily*, 2015, https://www.sciencedaily.com/releases/2015/06/150615124719.htm (last visiting 20/07/2018); R.E. Leenes, F. Lucivero, *Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design*, in *Law, Innovation and Technology*, 6(2), 2014, 193-220, doi: 10.5235/17579961.6.2.193; D. Tuffley, *Self-driving cars need 'adjustable ethics' set by owners*, in *The Conversation*, 2014, http://theconversation.com/self-driving-cars-need-adjustable-ethics-set-by-owners-30656 (last visited 20/07/2018); P. Lin, *The ethics of autonomous cars*, in *The Atlantic*, 2013, http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomouscars/280360 (last visited 20/07/2018); P. Lin, K. Abney, G.A. Bekey, *Robot ethics: The ethical and social implications of robotics*, Cambridge (MA), 2011; A. Wolkenstein (2018); G. Contissa *et al.* (2017).

[3] The aim of this paper is to address some issues that have the potential to enrich the concept of autonomy and provide materials to re-discuss this concept beyond a consequentialist or Kantian conception. In order to carry out this strategy, we will not adopt a culturally-oriented concept of autonomy. We do not want to make the mistake of harnessing traditional categories to a new problem. Instead, we will try to develop a new and specific framework of autonomy for machines. Our proposal claims not to be dependent on the technological level of complexity of the machines, as it is not a question of technical improvement, as much as a logical impossibility. This paper does not address the technicalities of A.I. systems, and therefore does not consider the present technical advancements on the matter as anything more than examples of automation. It does not aim to provide a mere overview of the current situation. Instead, it seeks to lay the groundwork of the matter as a whole. J. Achenbach, *Driverless cars are colliding with the creepy trolley problem*, in *The Washington Post*, 2015, https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-drivingcars-ever-solve-the-famous-and-creepy-trolley-problem/ (last visited 21/07/2017); A. Hern, *Self-driving cars don't care about your moral dilemmas*, in *The Guardian*, 2016, https://www.theguardian.com/technology/2016/aug/22/self-driving-cars-moral-dilemmas (last visited 20/07/2018).

[4] Are we sure that machines will share human values? This is a question that we do not try to solve in this paper. Anyway, since self-driving cars have been built with the aim of replacing humans, it is possibly interesting to make an analogy between man and machine for the decision-making process and the (moral) values involved, even though it is not clear whether human truths could be applied to machines.

[5] J.M. Paxton, J.D. Greene, *Moral reasoning: Hints and allegations*, in *Topics in Cognitive Science*, 2(3), 2010, 511-527, doi: 10.1111/j.1756-8765.2010.01096.x; P. Singer, *Ethics and intuitions*, in *The Journal of Ethics*, 9, 2005, 331-352, doi: 10.1007/s10892-005-3508-y; L. Petrinovich, P. O'Neill, M. Jorgensen, *An empirical study of moral intuitions: Towards an evolutionary ethics*, in *Journal of Personality and Social Psychology*, 64(3), 1993,:467-478.

will consider whether these two answers are founded on the structure of the human brain. We will then show the fallacy of the traditional 'one-right-answer' approach (the approach in which only one solution is undoubtedly right or wrong). Finally, we will explain our new proposal: the dilemmatic approach and its ability to give a better explanation of the problem of moral machines. This may also be a new way to consider improvements to A.I. (Artificial Intelligence) systems. A tragic choice may be considered as a new tool for enhancing the A.I. approach to human problems.

2. The trolley problem

Self-driving cars have the potential drastically to reduce the number of road accidents, but it is undoubtedly true to say that they will not be able to solve the problem tout court. This implies that there is a compelling necessity to consider scenarios in which autonomous vehicles would face problematic situations and to try to find adequate solutions.

In other words, the ethical and legal issues involving self-driving cars raise moral quandaries.

We are referring to conflicts involving a plurality of possible actions from among which the agent has to choose according to her/his 'moral' view[6].

From many experiments we can argue that it is a common sense solution to choose a moral principle and then fix the machine rule according to that principle. In doing this, we have to clarify what is considered as the best moral principle. Quoting Greene, «the problem, it seems, is more philosophical than technical. Before we can put our values into machines, we have to figure out how to make our values clear and consistent»[7].

Logically, this path seems to be as follows:

1. Machines will have our (human) moral values;
2. Before we put our values in them, we have to clarify those values; and
3. We need to make our human values reliable.

Thus, it seems very important to concentrate our attention first on the problem of making our values clear and consistent, from a 'human point of view'.

To discern the human point of view in the field of moral principles, we will use a particular kind of moral experiment, known as the trolley problem. This will help us to put aside the potential problematic aspects related to the specifics of every single one of the hundreds of possible individual quandaries, and to set our focus only on the details that actually serve a purpose in our reasoning.

The runaway trolley is a thought experiment, originally formulated by Philippa Ruth Foot in 1967[8]. The problem has been extensively analysed by Judith Jarvis Thomson[9] and Peter Unger[10], both of

[6] It is interesting to consider MIT's 'Moral Machine' experiment (http://moralmachine.mit.edu/ last visited 20/07/2018), which they describe as «a platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars». It may be argued that it is not clear whether the study aims to extend the question of morality to self-driving cars or simply to study human thought, morality and decision-making processes. It seems that the problem remains fully within the human dimension rather than focusing on the machine (as the title would suggest).

[7] J.D. Greene, *Our driverless dilemma*, in *Science*, 352(6293), 2016, 1515, doi: 10.1126/science.aaf9534.

[8] P. Foot, *The problem of abortion and the doctrine of the double effect*, in *Oxford Review*, 5, 1967, 5-15.

[9] J.J. Thomson, *Killing, letting die, and the trolley problem*, in *Monist*, 59(2), 1976, 204-217.

whom provided new variants to the original scenario, and has been subject to several modifications over time; we will now present two of the main formulations of the trolley problem, including Foot's original one.

a) First example: the runaway trolley (or bystander)

The runaway trolley problem is as follows. There is a runaway trolley hurtling down the railway tracks. Up ahead, you see five people on the track. They are tied up and unable to move, and the trolley is headed straight for them. If you do nothing, the five people will die. You are standing next to a lever, which can divert the trolley onto a side track, to which another person is tied up. You have two options: a) do nothing, letting the trolley kill five people on the main track, or b) pull the lever, letting the trolley kill one person on the side track. According to Judith Thomson's research, people answer the dilemma unanimously, yet the reasons for their answer differ greatly. Although everybody to whom she presented the trolley problem said that it was acceptable to divert the trolley onto the other track for a net saving of four lives, «some people say something stronger than that it is morally permissible for you to turn the trolley: They say that morally speaking, you must turn it – that morality requires you to do so. Others do not agree that morality requires you to turn the trolley, and even feel a certain discomfort at the idea of turning it»[11].

This dispute, however, does not lead to different results: if we intend the trolley problem as a choice between killing one person or killing five people, it seems justified that the vast majority find it acceptable to pull the lever and kill the one person to save the five[12].

Neuroscientific studies have reached the same conclusions as Thomson, finding that most test subjects said that pulling the lever was appropriate[13]. We should note that in Foot's original formulation the subject with the dilemmatic choice is the driver of the trolley, while in the bystander scenario the person who makes the choice is – as the name suggests – a bystander watching everything from a distance. We will take just the second scenario into consideration, as Greene's research[14] discriminates personal from impersonal situations, and the original formulation may cause greater (and, most of all, pointless) ambiguity, running the risk of letting the issue change from the already complicated neuroscientific explanation of the moral decision-making process to a non-essential discussion concerning objections connected to the unclear distinctive features between the two problems considered in the studies (i.e. the original trolley problem and the footbridge scenario, which we will now present).

b) Second example: the fat man (or footbridge)

The second famous representation of the trolley problem is the following: a trolley is barrelling down a track towards five people. You are on a bridge over the track and next to you stands a very fat man.

[10] P. Unger, *Living high and letting die: Our illusion of innocence*, New York, 1996.

[11] J.J. Thomson, *The trolley problem*, in *Yale Law Journal*, 94(6), 1985, 1395-1396.

[12] E.D. Nucci, *Self-sacrifice and the trolley problem*, in *Philosophical Psychology*, 26(5), 2013, 662-672, doi: 10.1080/09515089.2012.674664.

[13] J.D. Greene, R.B. Sommerville, L.E. Nystrom, J.M. Darley, J.D. Cohen, *An fMRI investigation of emotional engagement in moral judgment*, in *Science*, 293, 2001, 2105-2108, doi: 10.1126/science.1062872.

[14] J.D. Greene, *Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie*, in *Journal of Experimental Social Psychology*, 45(3), 2009, 581-584, doi: 10.1016/j.jesp.2009.01.003; J.D. Greene, R.B. Sommerville, L.E. Nystrom, J.M. Darley, J.D. Cohen, *An fMRI investigation of emotional engagement in moral judgment*, cit.

You can stop the trolley by pushing him onto the track. You have two options: a) do nothing, and let the trolley kill five people or b) kill the man to save the five.
In contrast to the reaction to the first example, the same majority of people who approve of pulling the lever to sacrifice one life and save five lives disapprove of pushing the man onto the track to save the same net number of four lives. This, according to Eric Rakowski, may be because killing somebody incidentally or indirectly in the course of saving others is allowable[15], «provided that the aggregate gains exceed the total losses by the proper amount», but pushing the man exceeds the margins of moral tolerability[16]. Similarly, Greene's results confirm Thomson's research and highlight how most test subjects consider that pushing the man is not appropriate[17].
In order to analyse this question properly, we will start by considering the two main moral inclinations that drive people who are faced with this moral problem: the utilitarian approach and the deontological approach[18]. The former implies that the morality of an action depends on its consequences. The latter, conversely, states that the morality of an action depends on its intrinsic nature.
Dual-process theories of moral judgment suggest that these two moral approaches guide the responses to moral problems. The dispute between the supporters of the two major ethical theories over moral supremacy is deeply rooted in a centuries-old philosophical debate.
The iconic representatives of the two theories are, respectively, Jeremy Bentham[19] and John Stuart Mill[20] for the utilitarian approach and Immanuel Kant[21] for the deontological approach.
There are, however, many interesting theories on the matter by twentieth-century exponents of both currents of thought, which can allow us to approach moral dilemmas with an eye for their concrete resolution. For this purpose, we will examine two twentieth-century philosophers, one being a supporter of utilitarianism and the other an advocate of deontology.

3. The Utilitarians. The Deontologists

As we have already mentioned, a utilitarian approach to morality implies that no moral rule or act is intrinsically right or wrong. Morality is therefore in no way an end in itself, but is a means to some other end.
Karl Popper in *The open society and its enemies* introduces a new kind of utilitarianism, which he calls 'negative utilitarianism'. In his work, Popper suggests that the utilitarian principle of maximizing

[15] J. Haidt, J. Baron, *Social roles and the moral judgement of acts and omissions*, in *European Journal of Social Psychology*, 26(2), 1996, 201-218, doi: 10.1002/(SICI)1099-0992(199603)26:2<201::AIDEJSP745>3.0.CO;2-J.
[16] E. Rakowski, *Taking and saving lives*, in *Columbia Law Review*, 1993, 1071.
[17] J.D. Greene, R.B. Sommerville, L.E. Nystrom, J.M. Darley, J.D. Cohen, *An fMRI investigation of emotional engagement in moral judgment*, *op. cit.*
[18] P. Conway, B. Gawronski, *Deontological and utilitarian inclinations in moral decision making: A process dissociation approach*, in *Journal of Personality and Social Psychology*, 104(2), 2013, 216-235, doi: 10.1037/a0031021.
[19] J. Bentham, *An introduction to the principles of morals and legislation*, Oxford, 1907 (Original work published 1789).
[20] J.S. Mill, *Utilitarianism*, New York, 1998 (Original work published 1861).
[21] I. Kant, *Groundwork of the metaphysics of morals*, Oxford, 2002 (Original work published 1785).

pleasure (the greatest good for the greatest number) should be replaced by a new principle: minimizing pain.

He argues that «there is, from the ethical point of view, no symmetry between suffering and happiness, or between pain and pleasure… Human suffering makes a direct moral appeal, namely, the appeal for help, while there is no similar call to increase the happiness of a man who is doing well anyway»[22].

According to Popper, people can avoid the risks of utopianism by acting to minimize suffering.

An example of this is the major issue that arises in various scientific papers recently published on the matter[23] concerns the opportunity for selfdriving cars to handle certain problematic situations by deciding to kill their passengers and thereby safeguarding the greater good[24].

Conversely, deontological ethics is based on three key features:

1. Acts and rules are intrinsically right or wrong;
2. People should be treated as subjects of intrinsic moral value; quoting Kant: «Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end»[25]; and
3. A moral principle is a categorical imperative and therefore must be applicable for everyone who is in the same moral situation.

Contemporary deontologist Frances Kamm[26], however, takes a slightly different view.

Kamm presents a principle (the principle of permissible harm) according to which one may harm in order to save more people if and only if the harm is an effect of an aspect of the greater good itself, not just an efficient means to prevent a greater evil.

The principle of permissible harm is an attempt to provide a deontological guideline for determining the circumstances in which subjects are permitted to cause harm to others with their actions.

Interestingly enough, the foundation of this form of deontology seems to be inclined towards a moderate consequential model.

Ultimately, by looking at these two examples of modern approaches to theories of ethics, it appears that the dividing line between the utilitarian and the deontological approach has narrowed enough

[22] K.R. Popper, *The open society and its enemies*, London, 1945, 284-285.

[23] J.D. Greene, F. Rossi, J. Tasioulas, K.B. Venable, B. Williams, *Embedding ethical principles in collective decision support systems*, in *30th AAAI Conference on Artificial Intelligence*, 2016, 4147-4151; G. Kahane, *The armchair and the trolley: An argument for experimental ethics*, in *Philosophical Studies*, 162(2), 2013, 421-445, doi: 10.1007/s11098-011-9775-5; D. Brutzman, D. Davis, G.R. Lucas, R. McGhee, *Run-time ethics checking for autonomous unmanned vehicles: Developing a practical approach*, in *Ethics*, 9(4), 2010, 357-383.

[24] J.F. Bonnefon, A. Shariff, I. Rahwan, *The social dilemma of autonomous vehicles*, in *Science*, 352(6293), 2016, 1573-1576, doi: 10.1126/science.aaf2654. On this point, Greene and colleagues highlight the necessity of a hybrid decision-making system (suitable both for humans and machines) following some form of «moral values and ethical principles, as well as safety constraints». They also point out that «humans would accept and trust more machines that behave as ethically as other humans in the same environment» (J.D. Greene, F. Rossi, J. Tasioulas, K.B. Venable, B. Williams, *Embedding ethical principles in collective decision support systems*, cit., 4147).

[25] I. Kant, *Groundwork of the metaphysics of morals*, cit., 429.

[26] F.M. Kamm, *Intricate ethics: Rights, responsibilities, and permissible harm*, Oxford, 2007.

to allow us to take into serious consideration the possibility of finding common ground between the two positions.

4. Neuro-explanation: how humans react

The possibilities of modern technology have allowed many and diverse interested parties to create their own trolley problems[27], submit them to the general public and receive feedback.
We will briefly present two examples of this.
The first test, published by the website *Philosophy experiments*, is called 'Should you kill the fat man?'[28].
The test presents a set of preliminary questions determining the subject's moral orientation. It then presents four classic scenarios (runaway trolley, fat man, fat villain, ticking bomb) and asks the subject to decide whether to pull the lever, push the man and torture the villain.
The aim of the test is to measure the subject's 'consistency score' (the average being 77%), which means the extent to which the subject's moral choices are governed by a small number of consistently applied moral principles. If this is not the case, the possibility arises for moral choices to be essentially arbitrary.
The second test is called 'Moral Machine'[29] and was created by Scalable Cooperation at the MIT Media Lab as part of a research study into ethics for autonomous cars.
The subjects are asked to judge random moral problems involving a driverless car that has to choose 'the lesser of two evils' (such as killing two passengers or two pedestrians).
The aim of the test is to encourage the public to reflect on important and complex decisions.
The relevant factors evaluated, and the average results from participants, are the following:

- saving more lives: does not matter / matters a lot (average: matters);
- protecting passengers: does not matter / matters a lot (average: indifferent);
- upholding the law: does not matter / matters a lot (average: matters);
- avoiding intervention: does not matter / matters a lot (average: indifferent);
- gender preference: male / female (average: indifferent);
- species preference: humans / pets (average: humans);
- age preference: younger / older (average: younger);

[27] It ought to be pointed out that the nature of this kind of test resembles that of a game. In both, the gamer/subject commits virtual acts that would not always be accepted by society. The moral significance of such acts, according to Rami Ali (R. Ali, *A new solution to the gamer's dilemma*, in *Ethics and Information Technology*, 17, 2015, 267-274, doi:10.1007/s10676-0159381-x), should be traced back to «the effect on the only moral agents involved in the act, the gamers and those observing their virtual acts» (R. Ali, *A new solution to the gamer's dilemma*, cit., 268). In this sense, we find ourselves in a sort of loop where the gamer is looking at society before committing virtual acts and society is observing. Society, however, is composed of single individuals, including the gamer of our example. This creates a sort of paradox, in the sense that – considering the case of the application of the trolley problem to autonomous vehicles, which are still relatively new – one asks society to take a stand, but what really happens is that people tend to base their reasoning on what observers may think in a context where they are both observed by society and part of the same society that observes.
[28] http://www.philosophyexperiments.com/fatman/Default.aspx (last visited 20/07/2018).
[29] http://moralmachine.mit.edu (last visited 20/07/2018).

- fitness preference: fit people / large people (average: indifferent);
- social value preference: higher / lower (average: higher).

The social reaction has not been without its studies. In fact, it has often been considered in the light of neuroscientific research. Thus, we ought to conduct a brief examination of this scenario from a new standpoint by relying on a different source, neuroscience, which has recently become more and more relevant within the scientific community and can provide us with an insight into moral decision-making processes[30].

In their paper *An fMRI investigation of emotional engagement in moral judgment*, Greene and colleagues[31] analyse the neurological processes related to moral decision-making in the two main trolley problem formulations (the bystander and footbridge scenarios).

In explaining the reasons for the different results, Greene *et al*. highlight three key factors[32]:

1. Whether the decision is linked to areas of the brain normally associated with emotions;
2. Whether the decision is personal or impersonal; and
3. Whether the decision is consistent with utilitarian or deontological moral approaches.

First of all, when analysing the role played by emotions, it is interesting that emotional areas are significantly more active in the footbridge dilemma than in the standard trolley scenario. The latter, conversely, involves increased activation in areas of the brain associated with cognitive processing[33].

The second step involves labelling the decision as personal or impersonal. The criteria applied by Greene et al. to distinguish between the two categories relate to the aforementioned greater or lesser response of emotional areas of the brain involved in the decision-making process.

Applying these standards, the standard trolley scenario shows more cognitive than emotional processes, and is therefore labelled as 'impersonal', while the footbridge scenario produces a greater involvement of emotional responses[34], thereby receiving the 'personal' label[35].

---

[30] F. Cushman, J.D. Greene, *Finding faults: How moral dilemmas illuminate cognitive structure*, in *Social Neuroscience*, 7(3), 2011, 269-79, doi: 10.1080/17470919.2011.614000.

[31] J.D. Greene, R.B. Sommerville, L.E. Nystrom, J.M. Darley, J.D. Cohen, *An fMRI investigation of emotional engagement in moral judgment*, cit.

[32] G. Kahane, N. Shackel, *Methodological issues in the neuroscience of moral judgment*, in *Mind and Language*, 25(5), 2010, 561-582, doi: 10.1111/j.1468-0017.2010.01401.x.

[33] E.B. Royzman, J.F. Landy, R.F. Leeman, *Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context*, in *Cognitive Science*, 39(2), 2015, 325-352, doi: 10.1111/cogs.12136.

[34] Greene's results seem to give partial confirmation of Nisbett and Wilson's study (R.E. Nisbett, T.D. Wilson, *Telling more than we can know: Verbal reports on mental processes*, in *Psychological Review*, 84(3), 1977, 231-259) on verbal reports on mental processes. These two authors make an assumption about the unconscious alterations of judgments by proposing a mechanism that drives people to make up 'reasonable-sounding justifications' for their choices when they are making them. Greene introduces a new important element: the result of various pieces of research conducted on the matter seem to acknowledge that people generally tend to remain aware of their «actual motives and subsequent rationalizations» (J.D. Greene, *The secret joke of Kant's soul*, in *Moral psychology,* in *Vol. 3: The neuroscience of morality: emotion, disease, and development*, Cambridge, 2007, 36) and therefore deliberately decide not to act in a rational way, preferring a different approach (G. Kahane, *On the wrong track: Process and content in moral psychology*, in *Mind and Language*, 27(5), 2012, 519-545, doi: 10.1111/mila.12001). On this point, see also R.E. Nisbett, T.D. Wilson, *The halo effect: Evidence for unconscious alteration of judgments*, in *Journal of Personality and Social Psychology*, 35(4), 1977, 250-256.

Lastly, Greene and colleagues try to link the preliminary results to a deontological or utilitarian moral approach.

In this way, the utilitarian outcome (pulling the lever and killing one person to save five others) has been described as consistent with impersonal and more cognitively-driven decisions[36]. Conversely, the deontological outcome (abstaining from pulling the lever) has been linked with personal and more strictly emotional decisions[37].

What the study reveals is a consistent correlation between certain areas of the brain and a specific approach towards a moral judgment[38], or, in other words, the concrete possibility that «different brain areas (emotional and cognitive) may control different types of moral reasoning (deontological and utilitarian)»[39].

According to Greene, «there is a substantial and growing body of evidence suggesting that much of what we do, we do unconsciously, and for reasons that are inaccessible to us»[40].

Neuroscientist David Eagleman[41] chooses a different method to evaluate people's responses to moral problems. He shows different videos – one for each formulation of the trolley problem – to test the subjects, asking them to watch the videos and subsequently within a narrow window of time (5 seconds) physically to pull a lever in front of them if they decide to make the trolley switch tracks (in the bystander scenario) or physically to push a doll to stop the trolley (in the footbridge scenario)[42].

In both cases, people are presented with the same problem: 'Would you trade one life for four?'.

Interestingly enough, all the participants decide to pull the lever in the bystander scenario, but do not push the doll in the footbridge scenario.

---

[35] J.D. Greene, F.A. Cushman, L.E. Stewart, K. Lowenberg, L.E. Nystrom, J.D. Cohen, *Pushing moral buttons: The interaction between personal force and intention in moral judgment*, in *Cognition*, 111(3), 2009, 364-371, doi: 10.1016/j.cognition.2009.02.001.

[36] It is interesting to take into consideration on this matter the study of the psychologist Jonathan Haidt on people's moral judgment (J. Haidt, *The new synthesis in moral psychology*, in *Science*, 316(5827), 2007, 998-1002, doi: 10.1126/science.1137651; J. Haidt, *The emotional dog and its rational tail: A social intuitionist approach to moral judgment*, in *Psychological Review*, 108(4), 2001, 814-834).

[37] According to Woods (A. K. Woods, *Moral judgments & international crimes: The disutility of desert*, in *Virginia Journal of International Law*, 52, 2012, 633-667), when test subjects feel 'an emotional surge' they tend to rely on moral heuristics ('do not harm'), while when they do not feel this surge they engage in utilitarian reasoning (J. May, *Moral judgment and deontology: Empirical developments*, in *Philosophy Compass*, 9(11), 2014, 745- 755, doi: 10.1111/phc3.12172; J. Moll, R. Oliveira-Souza, *Moral judgments, emotions, and the utilitarian brain*, in *Trends in Cognitive Sciences*, 11(8), 2007, 319-321, doi: 10.1016/j.tics.2007.06.001).

[38] J.D. Greene, *Moral tribes: Emotion, reason, and the gap between us and them*, New York, 2013; J.D. Greene, *Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains*, in *Trends in Cognitive Sciences*, 11(8), 2007, 322-323; J.D. Greene, L.E. Nystrom, A.D. Engell, J.M. Darley, J.D. Cohen, *The neural bases of cognitive conflict and control in moral judgment*, in *Neuron*, 44(2), 2004, 389-40; J.D. Greene, J. Haidt, *How (and where) does moral judgment work?*, in *Trends in Cognitive Sciences*, 6(12), 2002, 517-523.

[39] M.S. Pardo, D. Patterson, *Minds, brains, and law: The conceptual foundations of law and neuroscience,* Oxford, 2013, 56.

[40] J.D. Greene, *The secret joke of Kant's soul*, in *Moral psychology,* cit., 35.

[41] D. Eagleman, *The brain with David Eagleman*, on *PBS*, 2015, http://www.pbs.org/the-brain-with-davideagleman/home (aired between 14 October and 18 November 2015; last visited 21/07/2017).

[42] M. Sarlo, L. Lotto, A. Manfrinati, R. Rumiati, G. Gallicchio, D. Palomba, *Temporal dynamics of decision-making in moral dilemmas: An ERP study*, in *Journal of Cognitive Neuroscience*, 24(4), 2012, 1018-1029, doi: 10.1162/jocn_a_00146.

When asked about the reasons for their decisions, the participants seem to regard the first scenario as just a maths problem. The second scenario, on the other hand, implies an interaction between the subject and the man, with the former having to push the latter to his death. This aspect draws the line between what is morally appropriate and what is not, and it is in that very moment that the subject recruits other networks in the brain, those that are involved in emotion. Eagleman considers that every act of choosing is underpinned by a sort of neural conflict between our rational 'logical' self and our emotional self. This conflict is «born of tension between two systems in the brain»[43], the (logical) lateral networks and the (emotional) medial networks. Eagleman's results confirm the validity of Greene's distinction between personal and impersonal dilemmas that depends on the direct use of physical force and the subsequent predominant emotional response (or lack thereof).

These experiments are important for underlining the relationship between humans and moral problems[44]. As we have seen, humans are driven not only by reason: every choice is linked with the emotional sphere. Thus, human choices are not rational choices (if we take rational to be a synonym of logical)[45]. In addition to that, neuroscientific studies highlight the fact that socially optimal choices are made in an impersonal situation.

For these reasons, utilitarianism still seems the most effective alternative, at both a theoretical and a practical level. In addition, using Coleman and Holahan's explanation[46] of the difference between critical and conventional morality, the utilitarian choice seems the best conventionally moral solution.

5. The common solution to the trolley problem

In other words, considering the evolution of the trolley problem on a theoretical level and looking at the evidence provided by neuroscientists and more recently by the MIT study, it seems that finding an answer to it is not only possible, but is almost self-evident, in the sense that all the evidence points in the direction of a utilitarian approach to the issue.

[43] D. Eagleman, *The brain with David Eagleman*, cit., p. 4.

[44] M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin, J. Mikhail, *A dissociation between moral judgment and justification*, in *Mind and Language*, 22(1), 2007, 1-21.

[45] An example of this is that the majority of people seem to agree with the utilitarian moral solution: self-driving cars that are programmed to minimize the total amount of harm, even at the expense of their passengers. However, people's attitude toward such cars changes if another aspect comes into question: the perspective of travelling in them themselves. This could mean that, even though a utilitarian policy may seem the best choice to approach the problem at the theoretical level, this policy could not really be effective even if just a small percentage of the population actually used those vehicles. In this sense, if utilitarian cars were to be unpopular, the idea of increasing the number of autonomous vehicles to reduce road fatalities would not succeed. However, if a deontological approach was considered to be the best alternative, it is also clear that strong emotional intuitions may guide decision-makers, when facing problematic situations, to outcomes that do not maximize the utility.

[46] «In moral philosophy a distinction is drawn between principles of critical and conventional morality. Standards of conventional morality are authoritative if they are widely shared; standards of critical morality are authoritative not because they are shared but because they are correct or true», J.L. Coleman, W.L. Holahan, *Book review: Tragic choices*, in *California Law Review*, 67, 1979, 1380.

However, this position is not without its criticisms. What we mean is the following: the utilitarian solution seems to be a way not to solve the moral problem, but just to fill a gap.
The utilitarian solution seems to be an attempt not to solve the moral issue, but just to impose a utilitarian moral view on a social level, in the sense that, because they are constantly confronted with surveys and debates that reduce the trolley problem to a mere example of game theory, people are getting used to an utilitarian view of the problem, and its particular structure is becoming less and less relevant.
As an illustration, the MIT study[47] has been opened to the general public, and academics in the UK are trying to establish a committee to deal with problems of this type, having regard to society's beliefs, in order to provide a social licence for the technical side of the problem. It is more interesting to discuss how a choice is accepted rather than whether a machine may have the capability to make this choice on moral grounds. MIT's computational work is not a reason why the machine has the capacity to make an ethical decision, but it is in line with arguments about how a choice is accepted by humans.
We think that what is actually being done is to use machines as a means to build a utilitarian social ontology. It seems that society is being encouraged to take the utilitarian choice.
It is interesting to point out that MIT's Moral Machine can hardly be described as a non-partisan approach to the trolley problem, in the sense that the subject who has to make the decision is not presented with the possibility of following a proper deontological approach, since he or she has to decide who the machine is going to hurt and everything in the scenario stands still until he or she makes a choice. If, for instance, one were really to want to act in a deontological way (i.e. without intervening), it would be impossible to take part in the experiment, since one has to click physically on the best scenario in order to proceed. This prevents one from completely refraining from playing any sort of active role in such a situation, which constitutes the basis of the deontological position on the matter and is granted by other simulation games (for instance, the one retrievable on pippinbarr.com[48]).

6. Fallacies of the common approach

To sum up, whether we consider the utilitarian or the deontological option to be the best possible solution, we only reach a partial answer to the trolley problem. In order to find a more concrete answer, though, we ought to change our lens, since the one we have used so far seems to have brought us to a dead end.
What we want to highlight is the necessity to focus on a preliminary step that is often overlooked and taken for granted: instead of turning our attention to the search for a solution, we believe we ought to consider the 'tragic' aspect of the trolley problem and then we can claim that we are unable to solve the problem but can seriously face the dilemmatic situation.

[47] There is currently data analysis research being performed to understand how different cultures, political factions and so on would react to different choices.
[48] http://www.pippinbarr.com/games/trolleyproblem/TrolleyProblem.html(last visited 20/07/2018).

To cite Rittel and Webber[49], «because the process of solving the problem is identical with the process of understanding its nature, because there are no criteria for sufficient understanding and because there are no ends to the causal chains that link interacting open systems, the would-be planner can always try to do better»[50]. Also, «the planner terminates work on a wicked problem, not for reasons inherent in the 'logic' of the problem. He stops for considerations external to the problem: he runs out of time, or money, or patience. He finally says, 'that's good enough,' or 'this is the best thing I can do within the limitations of the project,' or 'I like this solution,' etc.»[51].

Reaching an understanding on how we should be allowed to solve the problem is therefore not the end of the circle, but just a step forward towards the best solution conceivable at a given time, in a given place.

In other words we have to deal with the so-called 'tragic choices'.

To quote Brown, «The missing dimension here is, I suggest, a sense of the tragic nature of the dilemmas we face – and perhaps of human existence itself»[52].

As stated by Michael Rasche, «at the same time, the phenomenon of tragedy is to be understood as a condition of philosophy, insofar the tragic is an expression of a mindset that allows philosophical thought»[53].

That is – we believe – where our society stands at the moment, and that is why there seems to be the need to find a solution to the trolley problem by the acquisition of what may be called a social licence[54]. According to Calabresi and Bobbitt, it is possible to learn about a society's moral standards by paying attention to the methods it uses when approaching tragic choices as well as the results to which it comes. The power of the tragic choice consists in forcing society to face the incompatibility between some of its norms, and consequently trying to overcome that incompatibility. The solution to these tragic choices remains, however, only temporary, as society is constantly changing.

On the other hand, though, we argue that, in order to be fully intelligent, we should not only be able to identify tragic situations, but also have the ability to comprehend them fully and overcome them.

7. Our proposal: a new approach

We think that, in trying to solve the trolley problem, the authors we have examined above take two things for granted:

1. They assume that this dilemma is – in fact – solvable;
2. They assume that there are only two possible solutions (the utilitarian one and the deontological one).

[49] H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, in *Policy Sciences*, 4, 1973, 155-169.
[50] H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, cit., 162.
[51] H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, cit., 162.
[52] C. Brown, *Tragic choices and contemporary international political theory*, in *SAGE Publications*, 21(1), 2007, 12, doi: 10.1177/0047117807073764.
[53] M. Rasche, *Der tragische Grund der Philosophie: Der Tragische als diachrone und synchrone Bedingung philosophischen Denkens*, in *Philosophisches Jahrbuch 2*, 2016, 317.
[54] G. Calabresi, P. Bobbitt, *Tragic choices*, New York, 1978.

With regard to the first issue, what we are trying to underline is the dilemmatic nature of the trolley problem, which very often ends up being dismissed as a negligible philosophical matter and therefore purely abstract.

The point is that what these approaches (the philosophical, neuroscientific, technological, and social approaches) lack is a clear and comprehensive picture of the matter. They seem to consider the tragic issue regarding the trolley problem as an obstacle that has to be overcome, when, in fact, the intrinsic dilemmatic nature of the problem makes it impossible to transcend this issue in the first place.

This is because every moral dilemma is in itself unsolvable by its very nature, or it would not be a dilemma. However, facing something that is unsolvable is not the same as not being able to choose.

On the contrary, it simply means that there is no correct choice in the matter. According to Rittel and Webber, one has to state a definitive formulation of the problem in order to be able to conceive all its possible solutions, but the formulation is the problem itself. They say: the information needed to understand the problem depends upon one's idea for solving it. That is to say: in order to describe a wicked-problem in sufficient detail, one has to develop an exhaustive inventory of all conceivable solutions ahead of time. The reason is that every question asking for additional information depends upon the understanding of the problem – and its resolution – at that time. Problem understanding and problem resolution are concomitant to each other. Therefore, in order to anticipate all questions (in order to anticipate all information required for resolution ahead of time), knowledge of all conceivable solutions is required[55].

We think that there is no difference between the procedure for formulating the problem and the search for solutions, since «every specification of the problem is a specification of the direction in which a treatment is considered»[56].

This is why the second statement (the assumption that there are only two possible solutions) becomes controversial.

In fact, the misconception here is the nature of the problem itself, which is based on a fundamental disagreement between two approaches whose origins can be traced back to generally accepted principles. The question, by contrast, seems to follow an aut-aut formulation.

That is to say, given these premises, the trolley problem is not a matter of choosing the preferable approach (whether deontological or utilitarian), but is instead the first step towards a real and far more complex dilemma.

By rejecting an aut-aut approach, we would like to point out something that does not come into the picture when discussing the trolley problem, which is that there have never been just two possible approaches to the matter: if we consider the problem in a strictly logical way, we see that there are four alternatives instead of two.

If we wished to represent the two approaches, deontological and utilitarian, in a logical form, we would do so by describing them as approach A and approach B, respectively. If we follow this line of thought, we cannot but consider their opposites as well, i.e. not A (-A) and not B (-B). At this point,

[55] H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, cit., 161.
[56] H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, cit., 161.

we ought to consider all the possible scenarios (every scenario in which A is and is not and B is and is not). We find four scenarios, each one describing one possible approach to the problem:
1) [A, B]; 2) [A, -B]; 3) [-A, B]; 4) [-A, -B]
It appears that the most interesting options are the ones that can provide a clear solution:
2) [A, -B] and 3) [-A, B]
In contrast to the typical modus operandi, we focus our attention on the two remaining alternatives:
1) [A, B] and 4) [-A, -B]
which do not offer a straight answer, and the latter of which, in particular, expresses the dilemmatic element within the trolley problem, the need to take into consideration the possibility of neither alternative being right.
We are referring above to the common aut-aut approach where the problem has to be answered in one way or another. In this sense, the possibility of both approaches being true [A, B] or both being false [-A, -B] is either not considered or is rejected. That is because the former would mean that the problem has no reason to exist, while the latter would not be able to provide a solution at all. This would mean that neither the deontological choice nor the utilitarian one was correct.
The implications of neither approach being right (in an absolute sense) are significant, on the grounds that the situation [-A, -B] does not lead to a dead end. On the contrary, stating that both A and B are not true suggests that the right answer may lie in a third approach that has not yet been considered.
Note that what we refer to as 'the right answer' must not be mistaken as the ultimate answer to the trolley problem as a mental experiment, since we do not believe that it is possible to solve this problem because of its dilemmatic nature. We only go so far as to suggest the possibility that there should be a better answer ('right' in a relative sense of the word) for each individual scenario, according to its peculiarities.
It would be common sense to suggest that this approach should lie in an indefinite point between A and B, and should be a more temperate version of both.
In this sense, Rittel and Webber propose a 'second generation' approach (as opposed to a 'first generation' approach that is adequate for classical problems), which «should be based on a model of planning as an argumentative process in the course of which an image of the problem and of the solution emerges gradually among the participants, as a product of incessant judgment, subjected to critical argument»[57].
From this point onwards, the path of our reasoning could be traced back to the same approach proposed by Abbott in *Flatland*[58].

[57] H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, cit., 162. As Rittel and Webber state in *Dilemmas in a general theory of planning*: «the classical paradigm of science and engineering is not applicable to the problem of open societal systems». That is because social problems are not solvable but instead, «at best they are only re-solved – over and over again» (H.W.J. Rittel, M.M. Webber, *Dilemmas in a general theory of planning*, cit., 160). Problems of this kind are therefore termed 'wicked' problems (while classical scientific problems are called 'tame' problems).
[58] In his novel, Abbott narrates the encounter between a square and a sphere, in which the latter teaches the former to go beyond its own perceptions and embrace the presence of several worlds, each with a different number of dimensions. At first, the square is reluctant to conceive of such a possibility, on the basis that it can-

In order to find the best possible approach for this specific context, we ought first to distance ourselves from the two opposing answers (granted that at this point none of them can be clearly identified as the best solution) and focus on their underlying principles.
That is to say, we acknowledge that neither approach, taken by itself, serves the purpose of deciding on a concrete plan of action that should be undertaken in a situation like the one depicted in the trolley problem. The problem lies in the isolation of the two conclusions: in order to analyse and compare them better, they have to collide. In such a way, the argument between the two positions will serve as a way to let the peculiarities of the two emerge and subsequently lead to a better identification of their common ground, as well as their discrepancies.
A comprehensive view of this sort may itself be enough to identify the best solution within this specific controversy. If not, we could operate a further breakdown of the elements of the evaluation and decide to consider directly the principles that constitute the basis of the specific principles of each position.

8. Conclusions

The implications of a process such as the one we have just described are significant: it is after levelling up from the rules to the underlying principles that we can find ourselves facing the most relevant alternative, or the choice between the non-decidable and the tragic.
The former would lead to a loop in the machine's calculus whenever a problematic scenario like the one depicted in the trolley problem presented itself in reality. In this situation, the matter would remain solely in the hands of humans, thereby relegating the machine to the submissive and purely mechanical position of a technical aid.
The latter, by contrast, would consist in a further deepening of the reasoning and a different definition of the rôle of the machine, which would only then have a real chance to become autonomous *tout court*. This is because: «[T]ragedy involves a situation where duties are in radical conflict, such that whatever is done will involve wrongdoing; by definition, this conflict cannot be wished away – the only way to preserve integrity and honour is to accept the tragic nature of one's choice: that is, to acknowledge that to act is to do wrong»[59].
In that regard, it is also proper to ask another question: since we specified in the premises of this work that we are considering the ways in which the decision-making process of AI can become closer and closer to the human decision-making process, we ought to ask ourselves how far we can look for similarities between the human and the artificial mind.
What does it mean for a machine to deal with the tragic?
There are – we think – two possible answers to this question.

not comprehend a three-dimensional world, which is too far from its reach as a consequence of its being a two-dimensional character. Later on, after visiting both the three-dimensional world and the one-dimensional world, not only does the square understand the various forms that reality can take, but it also starts to wonder whether reality could consist of an infinite number of dimensions, inaccessible to us, but present nonetheless. E. A. Abbott, *Flatland: A romance of many dimensions*, New York, 1963 (Original work published 1884).
[59] C. Brown, *Tragic choices and contemporary international political theory*, cit., 9.

On the one hand, one may claim that a moral machine can already face the tragic nature of a dilemma, since it is equipped to confront a dilemmatic situation.
On the other hand, in the present state of things and the present development of technology, machines can only go as far as recognizing the presence of a tragic dimension. This recognition leads irretrievably to a state commonly known as 'panic mode', following which a default solution may be activated.
In this sense, the moral dilemma remains within a human dimension and the machine is a mere executor of men's instructions and – therefore – of men's points of view and morality.
As part of this, since we are referring to the problem in light of the progress of artificial intelligence, we may also add a new category of assessments to our consideration, which are elements that constitute the machine's knowledge (derived from statistics, machine learning, etc.).
It is only at this point, we believe, that the machine's choice regarding the trolley dilemma would acquire the essentially autonomous nature that is ascribed to its definition as an autonomous machine and not just a slave.
We started from this point of view: it seems that people are trying to build moral machines with the aim of making them as close to humans as possible, and in order to do this scientists are trying to analyse human behaviour and thought processes and to replicate them.
If we really want machines to act like people, we should give them the chance to bear the sense of the tragic not only in a deadlock situation, but also in a situation in which one has to face a choice between two equally (desirable or) undesirable alternatives where making a decision results in a consequent deployment of conscience. It is well known that the tragic allows humans to experience their 'selves' (which is the aim of Ancient Greek theatrical tragedies). More specifically, we wonder whether dealing with the tragic could allow machines to experience the same self-discovery that is characteristic of this human experience.
Now it is interesting to consider, for example, self-driving vehicles and the meaning of this particular 'self'. If we consider 'self'-driving cars as vehicles that have the capability to drive autonomously, then we are talking about the machine as mere executor. By contrast, if 'self'-driving cars are seen as intelligent vehicles, there is a possibility for them to have consciences, and, if this is the case, then the paradigm will necessarily have to change, as a hypothetical trolley problem could neither be tackled nor solved on a purely human basis, but the new character should be considered as well.
We ought to highlight what we believe are the three possible meanings (there may be more) of the expression 'autonomous machine':

1. The machine is autonomous in the sense that, after the initial human programming, it can operate independently without the necessity of what is known as 'the man in the loop'. In this case, the machine is a mere executor of man's instructions;
2. The word autonomous may indicate that the machine is not tied or subordinated in any way to man's activity. The machine is fully independent;
3. Not only is the machine independent, but it also has a sense of morality.

Nowadays, machines are something onto which we download a series of rules, and efforts are being made for contemporary moral machines to go further. We suggest that more importance should be given to the capability of such machines to develop a sense of the tragic. Fully understanding the

sense and implications of a tragic situation does not mean that it is not possible to choose in favour of one of the possible solutions, but it just means that there are no longer right and wrong answers. What was once right or wrong is now better or worse, and this is where the two philosophical utilitarian and deontological principles become relevant.

That is to say, we know that facing and grasping the tragic nature of a problem allows a human being to elevate herself above the mere representation of the issue and, by investigating its premises, enables her to develop a sense of personality; can the same be said for machines?

Is it possible to imagine that, in presence of the tragic, a machine could rise above the mere (and inadequate) rules for the execution of a task by looking for a possible answer elsewhere, and by doing that could gain access to a deeper knowledge about the world and itself, to the point that if would be reasonable to consider the possibility of it developing a form of conscience?