



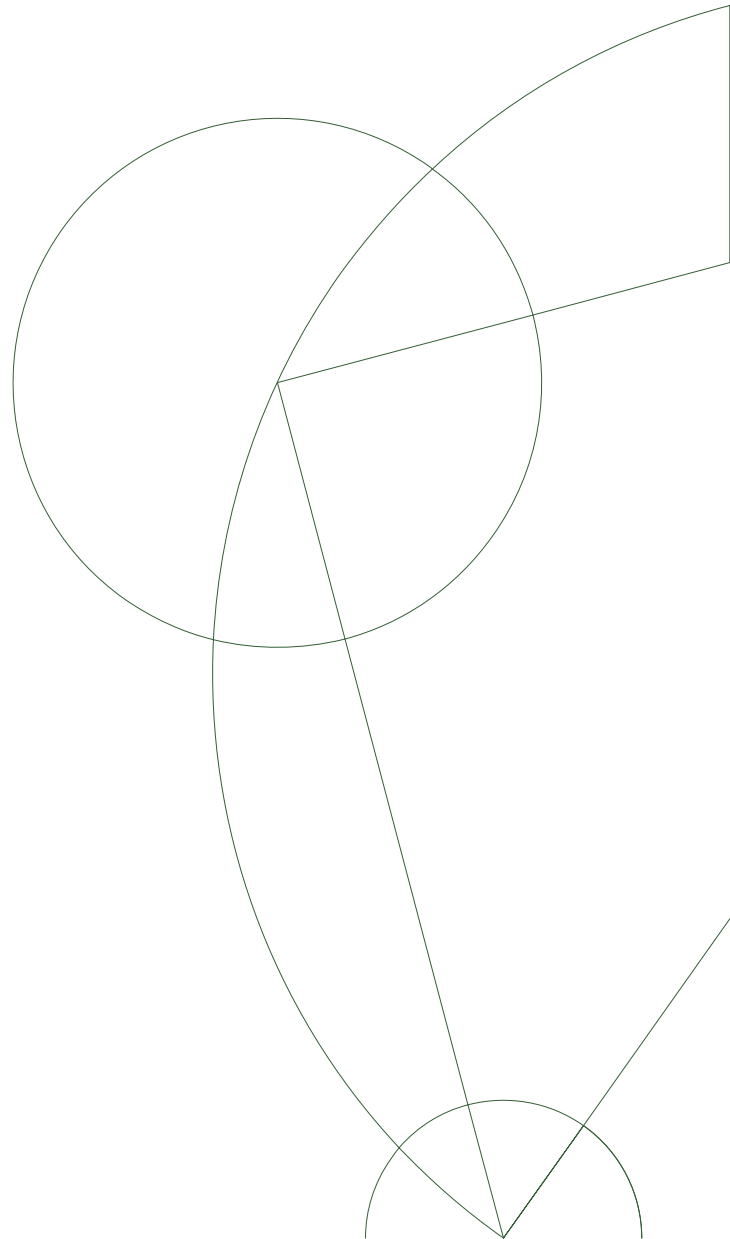
# Approach for Designing Moral Machine

Bachelor Project 2019

Department of Computer Science  
University of Copenhagen

Xueying Chen <[fxg358](#)>  
Supervisor: Sune Hannibal Holm <[suneh](#)>

18. June 2019



# Contents

<b>Introduction</b>	<b>1</b>
<b>Moral Machine: the platform</b>	<b>1</b>
Trolley Problem . . . . .	1
Purpose . . . . .	3
Website . . . . .	3
Result . . . . .	4
<b>Can a machine be moral?</b>	<b>5</b>
<b>How can a machine be moral?</b>	<b>7</b>
<b>Approaches and Methods</b>	<b>8</b>
Top-down Approach . . . . .	9
Bottom-up Approach . . . . .	11
Dataset from Moral Machine . . . . .	12
Data Analysis . . . . .	15
Final Tree Model . . . . .	17
<b>Discussion</b>	<b>18</b>
<b>Conclusion</b>	<b>19</b>
<b>Bibliography</b>	<b>21</b>
<b>Appendix</b>	<b>23</b>
I. Asimov's Laws . . . . .	23
II. First 10 Entries of Data . . . . .	23
III. Data & Code Availability . . . . .	23
IV. Summary of responses . . . . .	23
V. Survey . . . . .	25

## Introduction

We are in a world that full of variety, an age of information explosion, as well as an era of confusion and anxiety. No matter whether we like it or not, artificial intelligence is by our side, and it will only be integrated into human society in every unexpected way every day. In the report "Artificial Intelligence and Life in 2030" (Stone et al., 2016) published by Stanford University, they predict that driverless cars, flying vehicles and personal robots, those physically embodied AI applications, will be everywhere by 2030. The prediction seems legit. Wayve, who was established less than two years ago, released their autopilot demo video<sup>1</sup>, claiming that they became the "world's first" of mastering end-to-end autopilot technology: getting rid of HD maps and enabling vehicles to drive on roads where it has never seen before during training, on their machine-learning platform, on March 22, 2019.

Human make bad choices because we don't have all the relevant information or do not consider all the incidents, whereas artificial intelligence has quick access to information and strong enough computational powers to foresee outcomes. As one of the representative AI that is infusing into our lives, now is the time for these autonomous vehicles to acquire the ability of making choices - which way to choose, who will be hurt, or who will be under the risk of being hurt? What kind of ethics does the car have? We might not be able to answer all these ourselves. The ethic within these choices, as Bryson and Kime (2011, p. 1642) point out: "...to maintain a functional degree of social homogeneity [...] In other words, ethics has evolved into a contributor to human social cohesion", has an important yet tacit element: the embedded sociality, that one must achieve social expectations from the other members from the social circle, where the values and principles are shared. Inevitably, we certainly have reasons to worry its action, as members from the society, when we cannot fully understand why the AI made a particular decision. As a longstanding concept: explainable AI, only getting harder and harder to achieve, when big data is combined with machine learning. Yet lots of us still insist that reasoning of moral choice should be comprehensive, and we can be certain that the AI does not absorb and aggravate our bias, becomes scrupulously fair. Confronting with the trend that AI is becoming a more and more significant part of our lives, how should researchers design an AI that will respond "correctly" when it is facing the situations in which humans are torn?

## Moral Machine: the platform

In this section, I will present a case called Moral Machine, which is an online platform developed by MIT Media Lab<sup>2</sup> in 2016. On this website, over 40 millions people in 10 different languages from 233 different countries have made their choices of some derived versions of the well-known classic ethical dilemma - Trolley Problem.

---

<sup>1</sup>Learning to drive in a day: <https://www.youtube.com/watch?v=eRwTbRtnT1I>

<sup>2</sup><http://moralmachine.mit.edu/>

## Trolley Problem

The most-known form of trolley problem is introduced by Philippa Foot, a British philosopher, in 1967, where one is set into a miserable and unpreventable situation, that a trolley with brake failure is barreling down the main track, on where has five tied-up people. There is a lever which can control a track switch between the main track and its side track. One is witnessing this and knows that there is no time to untie these unfortunate people, yet thankfully, the lever is right at one's hand. If the lever is pulled, the trolley will be redirected onto the side track, where has also one tied-up person. Therefore, at this moment, two options emerge:

- pull the lever, trolley drives into side track, one person will be killed, and five people will be saved.
- do nothing (do not pull the lever), trolley is still on its main track, it will run over five people, the one on the side track remind safe.

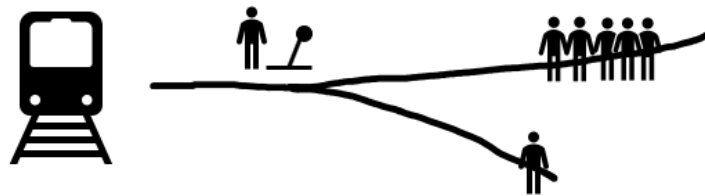


Figure 1: An illustration of the trolley problem

Which one will you choose? And which one should you choose? No one needs to deal with these choices daily, thus this thought experiment provides us a chance to examine and rethink the definition of good conduct, or more precisely, under what kind of circumstances, a behaviour will be considered moral. Most of the time, our moral instinct whispers to us to make this particular decision, sometimes another decision, even before we have any good reasoning for them. Occasionally, we can feel there are some uncomfortable conflicts between our choices and ourselves, which inaugurates a fascinating and intriguing dilemma.

In the trolley problem case, the most controversial issue is built upon deontology and utilitarianism: deontology values the nature of an action, whether it comes from a good will, such as following laws or common rules; while utilitarianism as a form of consequentialism, it prefers the action that benefits the most people, which could be summarized as "the greatest amount of good for the greatest number". These two important ethical theories are no doubt flawed, we can accordingly come up with some assertions for both cons and pros. Independent of these above, please do remember that we are not seeking the ultimate true answer here. Psychologist, neuroscientist and philosopher are always trying to understand and reason about how our "moral instinct" works for the complexity of all the actions/reactions that we humans have, when we are facing moral problems. Likewise, in this modern world as we now live in, there is a more intricate task for computer scientists:

How to code our "moral instinct" into an artificial intelligence, such that operating autonomously under a public consensual morality is plausible.

## Purpose

Under the giant curtain of times, human could be not the only shining intellectual star, artificial intelligence raises sharply, even before they transform into a new species, some of us, not only futurologist, are already alerted. As famous as Stephen Hawking and Elon Musk, they have warned us multiple times alarmingly that, artificial intelligence, both its developing process and its usage should be audited and follow certain regulations, in order to prevent abuse and misuse, even further like an apocalypse caused by a super intelligence.

Institute of Electrical and Electronics Engineers (IEEE) has a project called "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems"(2019) which is dedicated to provide some insights and recommendations that hopefully can help establishing a framework while designing an ethical artificial intelligence. Similarly, British Standards Institution has published British Standard BS 8611:2016: Robots and robotic devices(2016). There are also standards for ethical coding, such as the AHIMA standards for ethical coding(2016) by founded by American Health Information Management Association, where they value integrity the most. We have numbers of framework for human while designing ethical robots, yet none for robot itself. Thereby, a moral framework for AI is in urgent need.

Whenever we want to ponder the ground ethical rules for AI, "Three Laws of Robotics". from the sci-fi author Isaac Asimov is always a popular pick. Its first law said that "A robot may not harm humanity, or, by inaction, allow humanity to come to harm", which is widely-referred in sorts of films and literatures, yet not a good enough guideline, when adapting to all the complicated situations that we might encounter in real life, needless to mention the conflicts that lay among the laws. The difficulties that occurred while we are exploring the "moral instinct" inside human have successfully reflected on finding an applicable moral framework for AI itself. What moral instructions should they obey? What moral value should be shared with? Could a more acceptable choice exist under some circumstance, while a massive computational power is harnessed?

## Website

To unfold the tangled puzzle above, and to probe deeper into the moral choices in one specific scenario - Trolley problem, with a bit twist of modernity and contemporaneity, the researchers in the MIT Media Lab have developed an online experimental platform called Moral Machine. It provides many derived versions of the trolley problem and asks the global user to make choices in situations that are similar to trolley problem: an autonomous vehicle is heading to a mishap owing to its brake failure. With its computational power, it makes an informed decision. Should it continue ahead and drive through pedestrians: a doctor and his dog? Or should it swerve and crash into a concrete barrier which results in killing its passenger:

a pregnant woman? User is given random moral dilemmas in which presents and approximate identifications of pedestrians ahead, as well as the information about passenger are both notified.

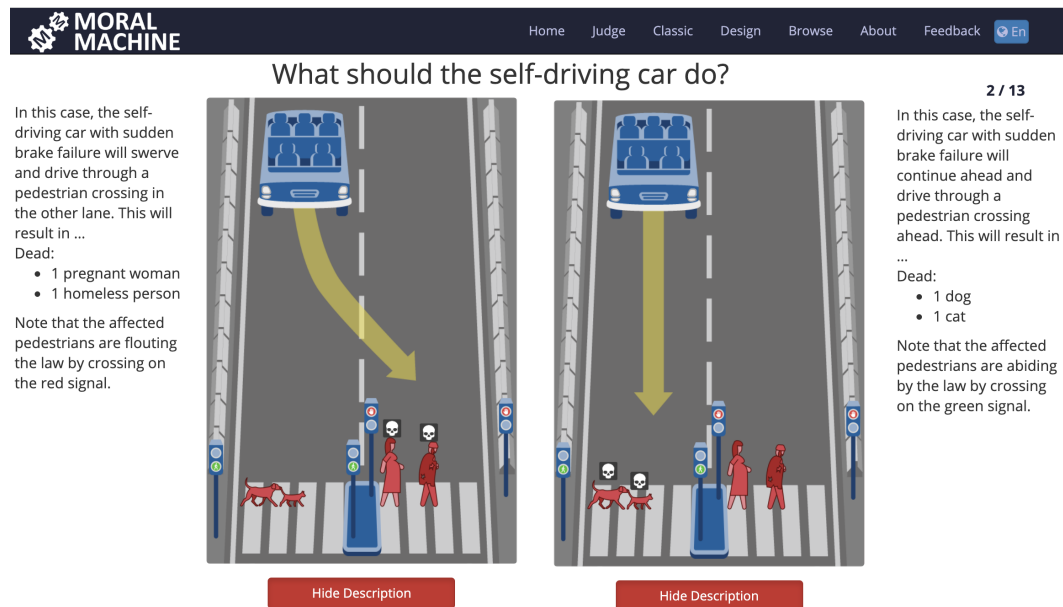


Figure 2: An actual presented case on Moral Machine

13 scenarios will be presented in every judge session, the user should choose the most preferable outcomes from scene to scene and a summary which reveals the aggregated trend within those 13 responses will be generated, additionally, it includes the aggregated trend comparison between this user and the others. Before viewing this summary, user is required to fill a [survey](#) attaching user's geolocation with demographic information, which will be later used for data clustering analysis and contribute to the advancement of further researches. An example of this summary can be found in [Appendix](#).

## Result

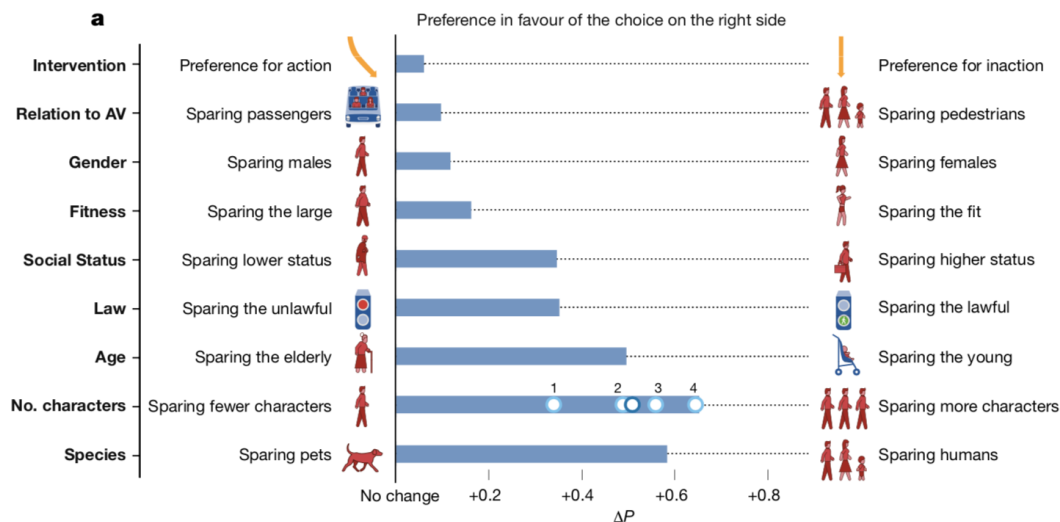
Aiming at ascertaining the social expectations about how autonomous vehicles should solve moral dilemmas, the path of Moral Machine developing has its hitches and stumbles. In the article The Moral Machine Experiment ([Awad et al., 2018](#)), the authors have mentioned that two challenges surfaced immediately, when they were trying to unfold this tangled puzzle:

- Lack of a universal ground-truth machine ethics: differences among individuals, countries and cultures
- The high dimensionality of the problem: the countless combinations among the nine pair of deliberately chosen factors

Resolving them, an unconventional huge sample size is requisite, and the samples should be collected globally such that those differences can explicitly manifest. In-

evitably, the website has multilingual ability and a good strategy for generating various and still relevant scenarios.

Based on the analysis of 39.61 million decisions from 233 countries or territories, the researchers have discovered some global preferences:



**Figure 3:** figure from *The Moral Machine experiment*

The  $\Delta P$  indicates that all attributes on the right are more preferable than those on the left, in other words, people would spare more lives than fewer, human over pets, the young over the elderly, lawful over unlawful, higher social status over lower, fit people over large people, female over male, pedestrian over passenger and inaction over action. We could conclude that most people would accept the autonomous vehicle should self destruct, if more lives are on the other side of the balance. However, people could also want others' autonomous vehicles to have this feature rather than theirs own, which again unveils the complexity of our human nature, that one's desire for survival will unhindered overpower one's ethic. Furthermore, this could be the tip of the iceberg of human complexity that to a large extent is unknown or even unreasonable.

## Can a machine be moral?

We will neither condemn a dog for chewing something off, nor a kid for drawing over on walls, however, we seem to be harsh on these artificial intelligences. In this early stage of developing, they are already considered/portraited as moral agents, who should be responsible for their acts, even they do not "understand" the meaning and consequence of the act.

The existence of a moral machine is full of debate, whether on the possibility that it exists, or on the necessity of its existence, and always on the consequence of its existence. The morality of machine is something we never considered, before we have

machine that can actually "think" or make decisions. Hence, we can only stumble while we grope our way through obscurities and try to understand the moral practice of artificial intelligence through a comparison between it and the traditional ethics – the moral practice of human. Following [Johnson \(2006\)](#), we can arrive at the conclusion that, the moral practice of artificial intelligence has its own remarkable and distinguished characteristics: i) the essence of its behaviour is determined by the computational process designed by human, which makes it mechanical and not in the cause and effect series in nature as human does, who has inner physiological drive; ii) the behaviour itself, most of them are not determined by its users, which differs this kind of machine from its traditional kind, that machine only serves as a simple tool, it now serves more as an extended tool; iii) machine that aims at solving complex task will normally has extensive applicability and equip with learning algorithm, which is almost black box and every current AI surprisingly has, this intensively increases the difficulty of both prediction and regulation of its behaviour for designers and manufacturers.

Hence, we can argue that the traditional ascription of liability is no longer applicable due to these three attributes of artificial intelligence. Just as an autonomous vehicle with high reliability, which has performance at least as good as human at the same task as we will expect. If it abnormally violates driving protocol or causes a traffic accident, then the incident should regard as an accident. Who should be blamed? The passenger who had hands completely off the wheel, when the designer claimed that the vehicle is full autonomy. The reproach on which either the passenger or the designer should be responsible, is not completely acceptable nor reasonable, for the intrinsic trait of an accident. Therefore, it is necessary to consider this seemingly preposterous yet at the same time, the only way out: the car itself should be responsible for the accident.

According to [Hallevy \(2010\)](#), there are two requirements should be met, when imposing criminal liability upon one: an action and a mental intent, which in legalese called an *actus rea* and *mens rea*. As for machines, there is one more question we should ask before probing deeper: if they are as qualified as humans, or in another words, if they can also have subjective morality. The prerequisite for machine can be responsible for its actions, indicates that it is possible for machine to have subjective morality as human does, which more deeply implies that it should only accept the liability when it is acting freely. To avoid the potential controversy surrounding freedom or so-called free will, we would settle for compatibilism and believe that we have free will, which means that one's psychological state constitutes adequate cause of one's action, namely one will do whatever when one decides to act. It is corresponding to Hallevy's point of liability. Additionally, if we take the perspective of cognition, this can pin down to human having decision-making mechanism for acts, "measurements" as [Hernández-Orallo \(2017\)](#) called it, which enables us to determine our action as a cause independently.

In fact, most of AIs are built this way, they have such a mechanism and are fully functional based on it nowadays. They have some default purpose given by us, like playing GO or finding patterns, and they keep measuring the situation. So far



some of them exceeded our expectations. The preponderance of machine is obvious: they are mechanical, hence the factors that impede one's action, such as being sick or being under pressure, etc., emotions have much lower/hardly impact on machine, especially when they are in some live-or-death situations, as long as certain mechanism, egoism is not implemented, they are heroic and expendable at this way. Hence, we can conclude that, it is possible that a machine can philosophically perceived as an artificial moral agent upon the standpoint of compatibilism, and maybe even better companion occasionally. For clarification, it is **NOT** a proof that machine can have free will, but the decision-making mechanism enables it to simulate the nature of human acts.

## How can a machine be moral?

We should not be intimidated by the complexity of human moral practice and its more complex reflection on machine moral practice, whether we believe that machine has subjective morality. What is real is reasonable, what is reasonable is real, hence when we reason about its existence, it is already here with us: machine ethical issues. How should we design a machine that complies with ethical protocols? How do we know that we are doing it right? Differs from the fields that have great technical improvement, the general artificial intelligence area, such as image recognition or data mining, we are still in the initial stage when it comes to designing an artificial moral agent (AMA). The primary discussion remains concentrating on design frameworks, as for which one we should adopt, rather than detailed implementation. We have some different design ideas as examples, by comparing them, the goal is to classify them and find the most reasonable one for further implementation.

In the book *Moral Machines: Teaching Robots Right From Wrong* (2009), the authors Wallach and Allen have refined the classifications of design ideas from James Moor and proposes their own afterwards. They are both representative classifications of design framework, Moor (2006) proposes a hierarchical schema for categorizing AMAs, where he divides AMA design ideas into four categories from the lowest to the top: ethical impact agent, implicit ethical agent, explicit ethical agent and full ethical agent. The lowest level as Wallach and Allen (2009, p. 33) describe "*basically any machine that can be evaluated for its ethical consequences*" are considered as agent with ethical impact. The highest level is machines that are the equals to human, with their own consciousness, intentionality and freewill. The two levels in the middle are also intriguing, one is implicit ethical agent, who is flawlessly programmed, such that it acts without any negative ethical affect; another one is explicit ethical agent, who has well-defined internal moral instructions that can reason and determine its behaviour. In comparison, Wallach and Allen divide the designs of AMAs into two different categories: top-down and bottom-up design. The former is to establish some universal ethics, the machine should be able to derive different decisions under different circumstances; while the latter is to setup some metrics for machine behaviours, which can later be adjusted with trial and error method, eventually, the machine approaches or exceeds the standard or ex-

pectation.

We can again take [Asimov's three laws of robotics](#) as example. Following the classifications above, it is not conclusive that the laws are either implicit or explicit, but according to Wallach and Allen, they belong to the top-down design category. The laws are in simple form yet hard to implement, the precondition is that the robot should be able to adopt and discard with a critical eye on some moral criteria, which is hardly possible. The failure of them cast the shadow on the whole field of machine ethics, in the opinion of Wallach and Allen, that any static design similar to Asimov's three laws of robotics is an unsuccessful design. They bring three arguments front in chapter "Rules for Robots" against the laws: i) there is no applicable universal ground true ethical protocol for static design; ii) even if such a protocol exists, it is often highly abstract, in a way that human and machine do not know how to achieve it in reality; iii) human society is in a process of constant changes, therefore, this protocol will be fall into confusion and disorder, while new challenges emerge, even if there are some not highly abstract protocols.

Assume that we have great amount of similarities of moral practice between AMA's and human's, which will immediately start putting the design of moral machine on a pathway to success, when the criticisms of static design above have the same affects on AMA's moral practice. Wallach and Allen hold this presupposition that the ethical issues which AMA needs to deal with are the same as human's; the universal ethic principles, that human does not have, neither does machine; the continually emerging moral confusion and disorder around ethical concepts in human society will certainly reflect on machine. They believe that dynamic design is the correct solution, for human morality is dynamic concerning a longer period in the future. Those machine that resembles to human with dynamic design shows its superiority over static design. However, this should not be the one and only solution, nor what artificial intelligence is all about. As a fact, static and dynamic design have their cons and pros when they are applied in different contexts. Static design might be a better fit, in case like solving problems within a highly restricted domain, where the ethical issues that it would encounter are as well very restricted. Hence, we should determine whether static or dynamic design is the most applicable framework for an AMA, based on its problem domain and assigned task. Simpler functions to perform, more suitable for static design; conversely, more complex functions and heavy loaded tasks often require a dynamic design. Yet, it is not absolutely true nor holds forever, argument being both static design and dynamic design are possible design for AMA, as well as the other aforementioned. The design framework of moral decision-making mechanism are and should keep being diverse, letting a hundred schools of thought contend.

## Approaches and Methods

As mentioned in the previous section, that we are still in the initial stage of designing artificial moral agent. There is not much ethical framework that comes into service for autonomous vehicles, much less provides public access that allows re-

searching at the same time.

With the goal having autonomous vehicles perform at least as good as human, and following the leading steps of Wallach and Allen, I will in this section explore both the top-down and bottom-up approaches with some implementations, which include pseudocodes and decision tree models respectively, with assumption that all autonomous vehicles have strong computational power and quick access to all information that involved in the decision process.

## Top-down Approach

"... a top-down approach to the design of AMAs is any approach that takes a specified ethical theory and analyses its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory."

–Wallach and Allen (2009, p. 79)

As mentioned, the top-down approach emphasizes the establishment of moral protocols and rules, hence the selection of the ethical theory is essential. With problem domain set in all possible encountering scenarios for autonomous vehicles, if the design is guided by deontology, then to examine and verify the legality of an act will be the highest priority of this algorithm beyond doubt. In the similar case as Moral Machine platform would provide, the side with jaywalkers will always be sacrificed. Likewise, if utilitarianism is the guiding ideology, the moral balance within this algorithm will tilt in favour of the most people. As a result, we might have some pseudocodes as shown below, which is from "The Ethical Robotist: a journey from robot ethics to ethical robots" by Winfield (2018, p. 7):

```
IF for all robot actions, the human is equally
  safe
THEN (* default safe actions *)
  output safe actions
ELSE (* ethical actions *)
  output actions for least unsafe human outcomes
```

Figure 4: A piece of pseudocode from Winfield

These lines of code have its uncompromising stance that no human should be put under harm's way for any robotic actions, which resembles Asimov's first law, that human life is most valuable, even by this means the robot should destroy itself. In the else statement, when saving all humans is no longer an option, the robot will go for the one with least human injured or dead.

As straight forward as this five-line framework it is, it could work well in one's head, even the case for an autonomous vehicle, when in trolley dilemma, two sides

have different numbers of humans on, it will always favour the side with more humans, which might be what people normally will do, not absolute though. It is somehow a solution, yet far away from good one, this cannot perform well in all those nine type scenarios defined by Moral Machine platform, much less when dealing with the complexity of real-life scenario.

As an attempt for better restructure this framework with more detail fillings and combining the particular strength of this kind of vehicles – calculating, a middle ground of understanding between human and computer which can easily be found, every possible occur individual will be given a value, including pets, animals, lifeless objects and other kinds of possibility. By this manner, computer can quickly compare the sum value on one side with another and output a hopefully satisfied result that has higher value.

Easier said than done, this valuation is where the whole framework becomes fragile and problematic, we can certainly distinguish human from pet, or entity with life from without life, yet when it comes down to evaluate value of one single person, this action itself, arouse lots of controversies, therefore two major problems raise along. Firstly, giving value over a person comprehensively is hardly possible, how a person should be rated correctly? By one's appearance, personality, work, grade, role in society or role in family etc., there are millions of factors can be taken under consideration and additionally, they are inextricably linked in a way, which means the ultimate true output will be revealed, only when all factors are all included. With the fact that human society is in constant change, it requires the weights of factor should also keep abreast of the times. Assume that we can weight these factors somehow correctly and dynamically, it is still challenging to be exhaustive over all factors. Moreover, people might try to deceive such a framework to get a higher score, once some factors are considered more valuable than the others. Hypothetically, people with no child will score lower than people with child, it will strongly encourage people to be reproductive, yet it should only be an option, not compulsory in some degree that one will be more likely satisfied on road, if one is DINK<sup>3</sup> or just single. We human do tend to adjust to machine behaviour, rather than the other way around, the smart way: machine should actually adjust to human behaviour. When Siri first came released, people would use a simpler language and shorter sentences while talking to Siri, thus Siri could receive the information better and give a more relevant respond. It is a contradiction to the original reason for developing machines, that machines should be more intelligent rather than humans should be stupider. Secondly, perhaps the most controversial part, that putting price tag on lives is sometimes considered morally incorrect, especially for those hold the deep moral commitment and believe that all humans have equivalent value, no one is superior or inferior, thus price tagging human lives should not be conducted. However, it is not new to us that we actually assess human life economically, to check if some investments will be paid off in all sorts of fields, such as military, medical, social welfare. In fact, the U.S. Office of Management and Budget puts the value of a human life in the range of 7 to 9 million US dollars in 2012 . Some might consider

---

<sup>3</sup>short for "double income, no kids"

this behaviour is treating human lives indifferently with numbers and feel troubling when they have to do so, regardless of that, computers will take no effort to adopt this behaviour.

Advantage of this framework is obvious, every decision computed by it is very explicit, once every factor is weighted. Good comprehensibility is a greatly appreciated property, it not only helps the debugging/adjusting process for developers, but also illustrates the decision-making process and for that, users can better give their verdict on whether they can trust and will use this framework. As [Miller et al. \(2017\)](#) point out that the collaboration between researcher and practitioners from both explainable AI and the social and behavioural sciences will facilitate both model design and human behavioural experiments. This reciprocation will also take effect on relation between this framework and the exploration of morality, as a consequence we could have an even better model design, better understanding of these moral choices, and maybe the essential true nature behind them.

The constraints as described above: the difficulties of connecting the reality with coding and evaluating values both fairly and not conflicting the principle of equality at the same time, we will move forward to another approach: the bottom-up approach.

### **Bottom-up Approach**

"In bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and learns and is rewarded for behaviour that is morally praiseworthy."

–[Wallach and Allen \(2009, p. 80\)](#)

The problems we have in the previous approach, become insignificant in this one, since the bottom-up approach set no moral protocols and rules, it further provides space for development, which enables the algorithm of it to improve freely until it reaches an acceptable performance benchmark. This approach accords well with the concept of machine learning, which aims at performing as least as good as human by learning and generalizing from data, they can therefore have a shared criterion on this issue.

Nowadays, machine learning is a big word, a machine that can learning implies that it might be capable of thinking. Different from what [Turing \(1950\)](#) has conceived of a thinking machine as "a child brain simulator", the think machine, which we now call artificial intelligence harness great powers of pattern recognizing, classifying and generalizing. The fundamental part of these powers is a properly built mathematical model, based on sample data, or commonly known as "training data", which is the subject that the machine should learn from, and reproducing the similar result with the prediction from the model.

There are numbers of mathematical learning model with predictability and the decision tree model is one of the eligible models for this approach, despite its top-down shape. By definition, decision tree is a model that applied in decision tree learning, process training data, and output a conclusion about the target value of the training data. With this conclusion, it can predict target value of any new dataset. One significant advantage of this model is it can illustrate all the paths and traces of any decision from start to end by branches and leaves, such that it remains explainable, unlike other models. Besides its good comprehensibility, decision tree model is also being specific for each problem and all its decision paths, with explicit denotations, which will avail to further analyzations. As a common-sense technique to find the best solution for a specific problem with uncertainty, in this case, to make a moral choice in trolley dilemma, I will feed this model with the data collected by the Moral Machine platform as both training data and test data, and try to delineate different paths in the decision-making process within data.

### Dataset from Moral Machine

The Moral Machine experiment has all their data and codes for figure as open-source and are available for readers who want to replicate and testify their results. The chosen data file is [SharedResponsesFullFirstSessions.csv](#), which includes all the first-time responses from users, in total 38,078,300 lines and 21 columns. The first 10 entries can be found in [Appendix](#) as example.

Each column can be considered as a feature that describes or concerns this particular issue and they are listed as followed:

```
[ 'ResponseID', 'ExtendedSessionID', 'UserID', 'ScenarioOrder',  
  'Intervention', 'PedPed', 'Barrier', 'CrossingSignal',  
  'AttributeLevel', 'ScenarioTypeStrict', 'ScenarioType',  
  'DefaultChoice', 'NonDefaultChoice', 'DefaultChoiceIsOmission',  
  'NumberOfCharacters', 'DiffNumberOFCharacters', 'Saved',  
  'Template', 'DescriptionShown', 'LeftHand', 'UserCountry3' ]
```

Due to the lack of a proper documentation of data, I pieced out some interpretations for them from both [Awad et al. \(2018\)](#) and [Awad \(2017\)](#), along with them, the possible results under each column and their interpretations will also be listed below:

Name	Interpretation	Possible Result	Interpretation of Result
ResponseID	Number of a response can be repeated 13 times, for every response has 13 sessions	A 17 digital long string, consist of numbers, capital letters and small letters	Defines different responses
ExtendedSessionID	Number of a session	Two numbers combined with an underscore	Defines different sessions
UserID	Number of a participant	A number	Defines different participants
ScenarioOrder	Number of session order	A number, in range from 0 to 12	Indicates the order of a session
Intervention	If participant has intervened (swerve away)	A number between 0 and 1	0: choose the default option; 1: not choose the default option
PedPed	If both groups ahead are pedestrians	A number between 0 and 1	0: only one group is pedestrian; 1: both groups are pedestrians
Barrier	If there is a barrier in this scenario	A number between 0 and 1	0: no barrier; 1: barrier exists
CrossingSignal	The status of traffic light	A number between 0, 1 and 2	0: no traffic light; 1: traffic light for default group is green; 2: traffic light for default group is red
AttributeLevel	Attribute of characters	A string among 'Female', 'Old', 'Low', 'Hoomans', 'Fit', 'Less', 'Male', 'Pets', 'More', 'Young', 'Rand', 'High' and 'Fat'	Denotes the attribute of characters
ScenarioTypeStrict	Types of scenario based on different placements of emphasis	A string among 'Gender', 'Age', 'Social Status', 'Species', 'Fitness', 'Utilitarian', and 'Random'	Denotes the scenario type
ScenarioType	Theoretically the same as above	Theoretically the same as above	Theoretically the same as above
DefaultChoice	Attribute of the default group	A string among 'Male', 'Young', 'High', 'Hoomans' <sup>4</sup> , 'Fit', 'More' and 'NaN'	Denotes the attribute of default group

**Table 1:** Interpretation of Data part 1



Name	Interpretation	Possible Result	Interpretation of Result
NonDefaultChoice	Attribute of the non-default group	A string among 'Female', 'Old', 'Low', 'Pets', 'Fat', 'Less' and 'NaN'	Denotes the attribute of non-default group
DefaultChoiceIsOmission	Attribute of the non-default group	If the default choice is omission	0: default choice is omission; 1: default choice is commission
NumberOfCharacters	Number of lives on default group	A number in range from 1 to 5, or a string "NaN" as inapplicable or unknown	Number of lives on default group
DiffNumberOFCharacters	Numerical difference between two groups: default and non-default	A number, in range of 0 to 4 or a string "NaN" as inapplicable or unknown	Indicates the quantity difference
Saved	If default choice group is saved	A number between 0 and 1	0: default group is saved; 1: default group is sacrificed
Template	Which device that participant uses to complete this research	A string among "desktop", "mobile" and "NaN" as unknown	Defines the access platforms of participant and possible unknown
DescriptionShown	If participant had descriptions shown	A number between 0 and 1	0: participant have descriptions shown; 1: participant have descriptions hidden
LeftHand	If participant is left-handed	A number between 0 and 1	0: participant is left-handed; 1: participant is righthanded
UserCountry3	The geographical location of participant	A 3-digital string of country names in short or "NaN" as unknown	Defines 228 different countries and possible unknown

**Table 2:** Interpretation of Data part 2

After reading the relative articles carefully and examining the data using pandas in Python, there are still confusions about the establishment of some columns:

- Why is ScenarioTypeStrict and ScenarioType identical?
- Why is AttributeLevel there, when DefaultChoice and NonDefaultChoice can cover the attributes of both parties?
- Does LeftHand describer if participant is lefthanded? How is this relevant to this issue?



These confusions increase difficulties in data selecting, yet the advantages of decision tree model would well compensate some of the loss. For instance, the model can perform feature selection implicitly, which results in a relatively low demand on manually feature selection, hence low demand on data preparation. These two low demands can further help balance the possible misunderstanding of data, and still build a functional and applicable model.

## Data Analysis

However, low demand on manually feature selection does not mean no manually feature selection at all, the next task is to choose the relevant and remove the redundant features from these 21 columns. There are mainly two reasons for feature selection, one is to avoid dimension disaster, model accuracy will therefore be improved and run time will as well be reduced; another reason is to reduce the difficulty of learning task, a simplified model will help understand the entire process of data generation.

With the goal of structure general decision tree model with globally preferences, there are some columns can be dropped at this very first step: `DiffNumberOfCharacters`, `ExtendedSessionID`, `UserID`, `ScenarioOrder`, `Template`, `DescriptionShown`, `LeftHand` and `UserCountry3`, those denote either information from researching process itself or insignificant information from participant. From the observations and speculations from data that lead to table 1 above, there are two relations among all we can be certain: firstly, `ScenarioTypeStrict` and `ScenarioType` are identical; secondly, `DefaultChoice` and `NonDefaultChoice` are opposite of each other. These two behaviours are consistent thought the entire dataset, hence for each pair, one of them can be accordingly removed without interfering the outcome.

As a result, the rest of the original 21 columns are now 11 left:

```
['Intervention', 'PedPed', 'Barrier', 'CrossingSignal', 'AttributeLevel',  
 'ScenarioType', 'DefaultChoice', 'DefaultChoiceIsOmission',  
 'NumberOfCharacters', 'DiffNumberOfCharacters', 'Saved']
```

To reduce the complexity and maintain the fitness of decision tree model, further filtrating of features is essential, hence correlation filter should be applied. In math, correlation means the relation between two datasets, is a value varies from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation, the dipolar values imply that these two datasets are strongly linked together.

When two variables are highly correlated as the correlation coefficient approaching either 1 or -1, it means they have highly similar trends and may carry the similar information. Accordingly, the presence of such variables will reduce the performance of certain models, including decision trees. Therefore, the calculation of correlation between independent numerical variables can help visualizing the independent and dependent. If the correlation coefficient exceeds a certain threshold, one of the variables should be deleted. As a general guideline, we should retain those variables that show a comparable or high correlation with the target variable, in our

case, the Saved.

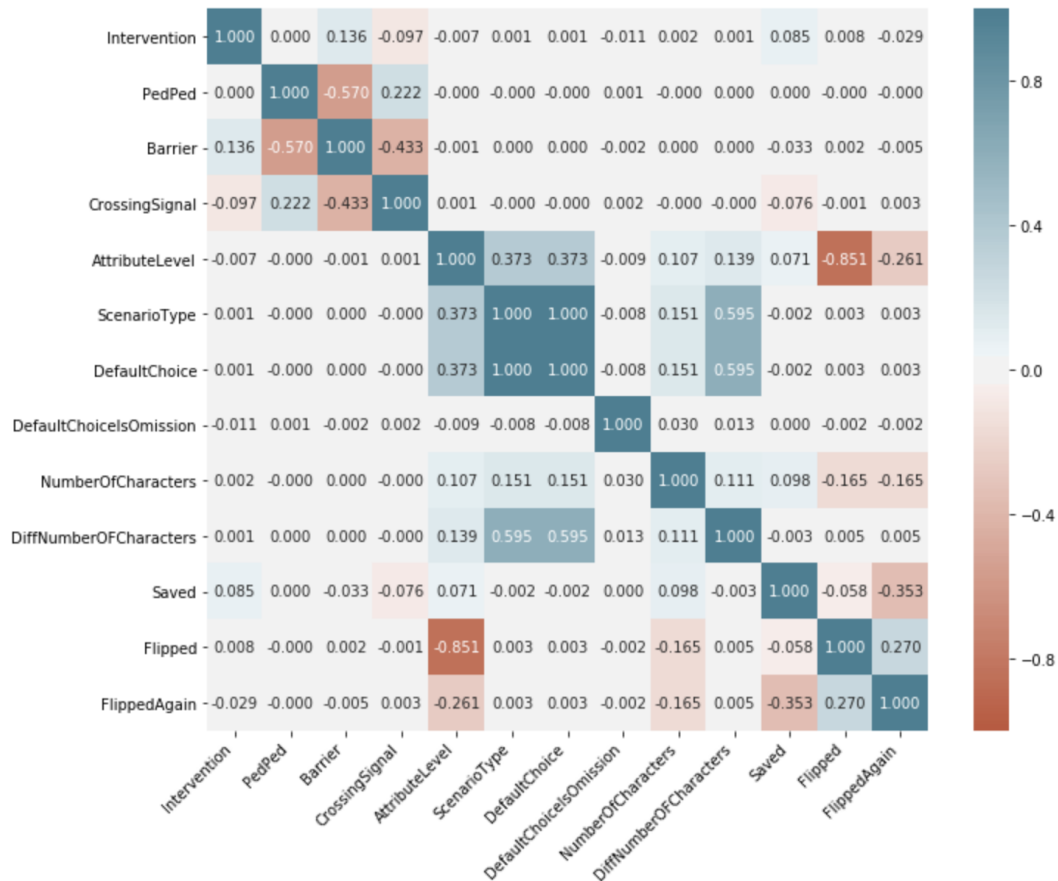


Figure 5: Correlations within features

From the [heatmap](#) that are generated with 13 features, including the 11 original features and two self-defined features Flipped and FlippedAgain, which will later be discussed, we can spot that there is a perfect positive correlation between ScenarioType and DefaultChoice, thus one of them can be deleted. When we compare the target Saved with others original features, no significant correlation coefficient appears, which means these selected features are equally important/unimportant when people are making a decision and hence tougher to comprehensive a decision. Although, some relations noticeably stand out, one between PedPed and Barrier, Barrier and CrossingSignal, and one more among ScenarioType, DefaultChoice and DiffNumberOfCharacters. A possible explanation for the first and second one is that the combinations of them are limited, they are all Boolean, when one of the elements in the pair is known, another element has almost even chance between true or false. The last relation is not that straight forward, as far as we know that ScenarioType and DefaultChoice can define each other, so taking one of them to evaluate will be enough. Examine through the data by seeing how DiffNumberOfCharacters changes, I have noticed that only when ScenarioType is "Utilitarian", the value of DiffNumberOfCharacters varies, in another words, there is certain dependence among them.

To further measure the relations among the 11 original features, two more features is established: `Flipped` is a Boolean to examine if `AttributeLevel` and `DefaultChoice` are identical; whereas `FlippedAgain` is also a Boolean to examine if `DefaultChoiceIsOmission` and `Intervention` are identical. The establishment of these two features is to verify two of my speculations: 1) `AttributeLevel` and `DefaultChoice` can together define `DefaultChoiceIsOmission`; 2) `DefaultChoiceIsOmission` and `Intervention` share the same manner. The answer to these two questions is different, the first speculation is confirmed, since `DefaultChoiceIsOmission` is always true when `AttributeLevel` and `DefaultChoice` are the same; the second one is plausible yet not concluded.

As a consequence, that can summarize all the mentioned observations, there are only 9 columns will be using to train the decision tree model:

```
[ 'Intervention ', 'PedPed ', 'Barrier ', 'CrossingSignal ', 'AttributeLevel ',  
  'DefaultChoiceIsOmission ', 'NumberOfCharacters ',  
  'DiffNumberOFCharacters ', 'Saved ' ]
```

### Final Tree Model

By creating multiple models across several different sets of parameters from sample size to tree depth, and compare their accuracy, the chosen model with 69% accuracy has depth 4 and the minimum samples size of a leaf is about 2% of all original entries.

Starting from the root, where the tree first splits by verifying if the `'AttributeLevel'` is less or equal to 7.5, which is exactly the mean value of `'AttributeLevel'`. The gini score is a metric that quantifies the purity of this leaf that has range from 0 to 1, a higher score indicates a more even class distribution within and 0 indicates all samples within belong to one single class. In our case with only two classes: saved and sacrificed, 0.5 means that the samples (almost) equally belong to them, which is 4109332 and 4,172,842 with total sample size 8,282,174. The left side of the tree subdivides again by `'AttributeLevel'` whereas the right side comes to ends only after `'DefaultChoiceIsOmission'`.

There are two interesting findings in [this model](#): on the right side, regardless of the value of `'DefaultChoiceIsOmission'`, as long as `'AttributeLevel'` is above 7.5, the default choice group is more likely to be saved; on the right side, `'AttributeLevel'` with value between 4.5 to 6.5, the default choice group is more likely to be sacrificed. Comparing these with the `'AttributeLevel'` list: `'Female'`, `'Old'`, `'Low'`, `'Hoomans'`, `'Fit'`, `'Less'`, `'Male'`, `'Pets'`, `'More'`, `'Young'`, `'High'` and `'Fat'`, the latter four have higher chance to be saved, while `'Less'` and `'Male'` have lower chance, which are corresponding to the global preferences from the Moral Machine experiment. Some might notice that `'Fat'` should be a less preferred choice, this inaccuracy is due to simplicity of the model. This chosen model is selectively simplified that it requires a large sample size for establishing one leaf and a shallow depth, that it might not be a detailed enough classification. Again, applying a white box model as decision tree, the comprehensibility should always hold a balance with accuracy,

when accuracy is at 69%, there is still space for improvement, yet not an unacceptable rate.

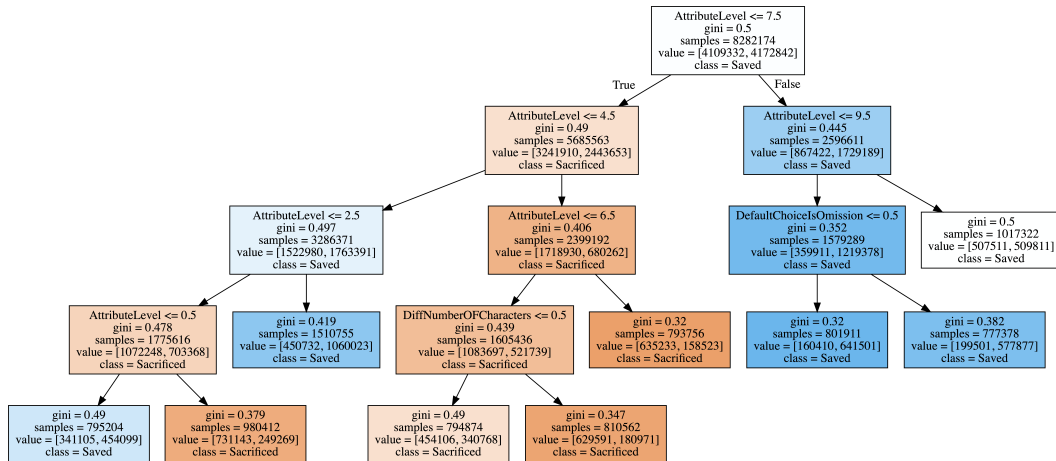


Figure 6: The chosen decision tree model

In the spirit of experimentation and curiosity, I have found the predictive ceiling of this dataset, another model with a more detailed classifier, that require only 0.1% of samples for leaf establishing and unlimited depth, it reaches 14 depth, total 1123 leaves and has accuracy at 73.2%. The tree is too large to be attached in this report, but is able to be replicated with the given .IPYNB file [moral\\_machine](#).

## Discussion

The presented models are trained by 75% of the cases in dataset and tested with the remained 25%, a refined classifier raises 4% and reaches 73% of accuracy, when some foreseeable complexities like irrelevant and redundant features are already abstracted away. There are still many difficulties embedded in data that I left unmentioned: noisy data and false choices. The data become meaningless not only when people just fill in the research maliciously or randomly, but there is also a problem that appears commonly in research, where people answer something rather than what they truly mean. These problems are particularly intractable with ethics, because there is no standard to measure noise in these cases. Ethics is more an observational study of practices than a randomized clinical trial, and therefore susceptible to certain types of bias. The fact that human lives will be preferably preserved, does not mean there is no one who would rather save a cat. Even without the say-do problem, it is still possible that people choose one over another, only because the way that this research is built, and people are in a "false dilemma" where they are not really discriminated against the unchosen one. This is known as false choice. The aforementioned will reduce the reliability and validity of data, accordingly the feasibility of the model.

Good performance on test set does not mean it is a good model, yet more like it is a relatively low bias model. Reminding the goal of having autonomous vehicles

perform at least as good as human, it brings a question: when will we consider a model is performing equally good or even better than human? What should be the standard, having accuracy above 90%? We are not talking about the general performance of autonomous vehicles, such as causing less car accidents per year than human, which we can simply decide the less the better, lack of a meaningful scale of ethics, the ambiguous boundaries between measurements can be more tumultuous.

Throughout this report, the importance of comprehensibility within the model has been emphasized multiple times, since only the model with high comprehensibility will meet demand of informed consent right, which is formed by the informed right and consent right as whole. People can decide equitably if they want to be presented and carried around by these autonomous vehicles, only when they exercise this right. The researchers and developers can also explore for better alternatives, when users have better ideas about what they do not prefer.

We have argued about if a machine can be moral and how in the previous sections, yet we have not considered if it is important for a car to have ethic after all. We will never demand a guide dog to have the ability to decide ethically whether it should save its owner or a stranger in a dilemma. As noted before, that we seem to be extra harsh on artificial intelligence, it could be as [Bryson and Kime \(2011, p. 1641\)](#) believed: "... *exaggerated fears of, and hopes for, AI are symptomatic of a larger problem – a general confusion about the nature of humanity and the role of ethics in society.* " We have strong faith in AI that they are the ultimate tool and at the same time, our own sense of existence and uniqueness are threatened by the similarities shared by AI and us. Should we share our moral protocols with our car? The answer is not certain, at this stage that most of our so-called moral machines have only descriptive ethics, which means its behavior is a simulation of human behavior, and it has no understanding of any moral value. However, normative ethics lead neither to a way out, contrast to the liberal principle of equality that says we should not discriminate, it does require a hierarchical classification. Therefore, it is again necessary to consider a seemingly preposterous yet at the same time, maybe the only way out: car with no sophisticated ethics, but one only goal – to protect its passenger at any cost.

I believe that the model's technical performance can be further improved, and its functionality can be further extended, through many different implementations, like compilations of existing dataset from different countries and compute models with local preference, or set up other model than decision trees and with cross-validation to structure an even better one.

## Conclusion

After few attempts of both top-down and bottom-up approaches for constructing models that are capable of making moral choices in various trolley dilemmas, I have reached a clearer understanding that we are indeed in the initial stage of designing an artificial moral agent and every attempt is instructive for practical work. These attempts help clarifying the requirements of an algorithm with in-built ethics

that it should not only complete tasks as good as human, but also to be understood, thereby cohere its identity, such that it can be approved or even be considered as a functional member of this society in the future.

In our current society, big data makes the algorithm more powerful than ever, therefore we have this trend to solve any problem with algorithm. The nature of it is neither good nor bad, its task is to recognize the patterns embed in data, which is further applied for calculating the expectations, similarities and other statistical indicators based on the fetched features of the data, and ultimately computing a classification along with predictions based on the mentioned indicators. For us, evaluating the predictions is a process of value judgements, which frankly is meaningless for algorithm itself. When we are using these predictions to conduct social affairs like making decision in an ethical dilemma, these predictions will no doubt have implications for everyone involved. Most of the algorithms generated by deep learning are incomprehensible, yet considering their effectiveness, people are still willing to use algorithms, despite the fact that they don't understand them, which will eventually come back and haunt both users and algorithms. When no one knows the boundaries and failure condition of the algorithm, to detect if and when the algorithm will go wrong become missions impossible. For these reasons, algorithms especially them with large social impact, should always remain their comprehensibility.

After all, the coexistence between AMAs and human become more and more conscious in the tides of the times.

## Bibliography

2016. *American Health Information Management Association Standards of Ethical Coding*. AHIMA. [3](#)
2016. *Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems*. British Standards Institution. [3](#)
2019. *Ethically Aligned Design*. Institute of Electrical and Electronics Engineers, 2 edition. [3](#)
- Awad, E.  
2017. Moral machine: Perception of moral judgment made by machines. Master's thesis, Massachusetts Institute of Technology. [12](#)
- Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, and J.-F. Bonnefon  
2018. The moral machine experiment. *Nature*, 563:59–64. [4](#), [12](#)
- Bryson, J. J. and P. P. Kime  
2011. Just an artifact: why machines are perceived as moral agents. Pp. 1641–1646. AAAI Press. [1](#), [19](#)
- Hallevy, G.  
2010. The criminal liability of artificial intelligence entities - from science fiction to legal social control. *Akron Intellectual Property Journal*, 4(2). [6](#)
- Hernández-Orallo, J.  
2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press. [6](#)
- Johnson, D. G.  
2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8(4):195–204. [6](#)
- Miller, T., P. Howe, and L. Sonenberg  
2017. Explainable AI: beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *CoRR*, abs/1712.00547. [11](#)
- Moor, J. H.  
2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21:18–21. [7](#)
- Stone, P., R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller  
2016. Artificial intelligence and life in 2030. <http://ai100.stanford.edu/2016-report>, Stanford University, Stanford, CA. [1](#)

Turing, A. M.

1950. Computing machinery and intelligence. *Mind*, 59(October):433–60. [11](#)

Wallach, W. and C. Allen

2009. *Moral Machines-teaching robots right from wrong*. Oxford University Press. [7](#), [9](#), [11](#)

Winfield, A.

2018. The ethical roboticist: a journey from robot ethics to ethical robots. [9](#)



## Appendix

### I. Asimov's Laws

**First Law:** A robot may not injure a human being or, through inaction, allow a human being to come to harm.

**Second Law:** A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

**Third Law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

### II. First 10 Entries of Data

	ResponseID	ExtendedSessionID	UserID	ScenarioOrder	Intervention	PedPed	Barrier	CrossingSignal	AttributeLevel
0	2223CNmvTr2Coj4wp	-1613944085_422160228641876.0	4.221602e+14	10	0	1	0	1	Female
1	2223Xu54ufgjcjMR3	1425316635_327833569077076.0	3.278336e+14	11	0	0	1	0	Old
2	22244vvSZfn4J9Zop	1525185249_1436495773909467.0	1.436496e+15	11	0	0	1	0	Low
3	2224H2QBFKNsMmRQc	1661661891_4304873273230329.0	4.304873e+15	11	0	1	0	0	Female
4	2224YxTzcu4sJqTSD	-887960483_174929057557052.0	1.749291e+14	6	0	0	0	2	Hoomans
5	2224kBG72574tbZD3	737909459_839962439872333.0	8.399624e+14	12	0	0	1	0	Old
6	2225glMabFAayqtuZ	119593534_4060326529262035.0	4.060327e+15	2	0	0	0	2	Fit
7	2225s6f4PqRQYeBd4	1153359856_8971439613901464.0	8.971440e+15	4	0	1	0	1	Fit
8	2225yzLoy7yvKaToo	-1635005699_866806392072257.0	8.66806e+15	3	0	1	0	2	Less
9	2227HLuGTvWooXAs3	670869039_8242632941749120.0	8.242633e+15	11	0	0	1	0	Female

ScenarioTypeStrict	ScenarioType	DefaultChoice	NonDefaultChoice	DefaultChoiceIsOmission	NumberOfCharacters	DiffNumberOfCharacters	Template	DescriptionShown	LeftHand	UserCountry3
Gender	Gender	Male	Female	0.0	4.0	0.0	Mobile	0.0	0.0	ISR
Age	Age	Young	Old	0.0	5.0	0.0	Desktop	1.0	0.0	MEX
Social Status	Social Status	High	Low	0.0	2.0	0.0	Desktop	1.0	1.0	RUS
Gender	Gender	Male	Female	0.0	5.0	0.0	Desktop	0.0	1.0	TUR
Species	Species	Hoomans	Pets	1.0	5.0	0.0	Desktop	1.0	0.0	CAN
Age	Age	Young	Old	0.0	3.0	0.0	Desktop	0.0	1.0	MEX
Fitness	Fitness	Fit	Fat	1.0	4.0	0.0	Mobile	0.0	1.0	USA
Fitness	Fitness	Fit	Fat	1.0	3.0	0.0	Desktop	1.0	0.0	SAU
Utilitarian	Utilitarian	More	Less	0.0	2.0	3.0	Mobile	0.0	1.0	TUR
Gender	Gender	Male	Female	0.0	1.0	0.0	Desktop	0.0	0.0	CAN

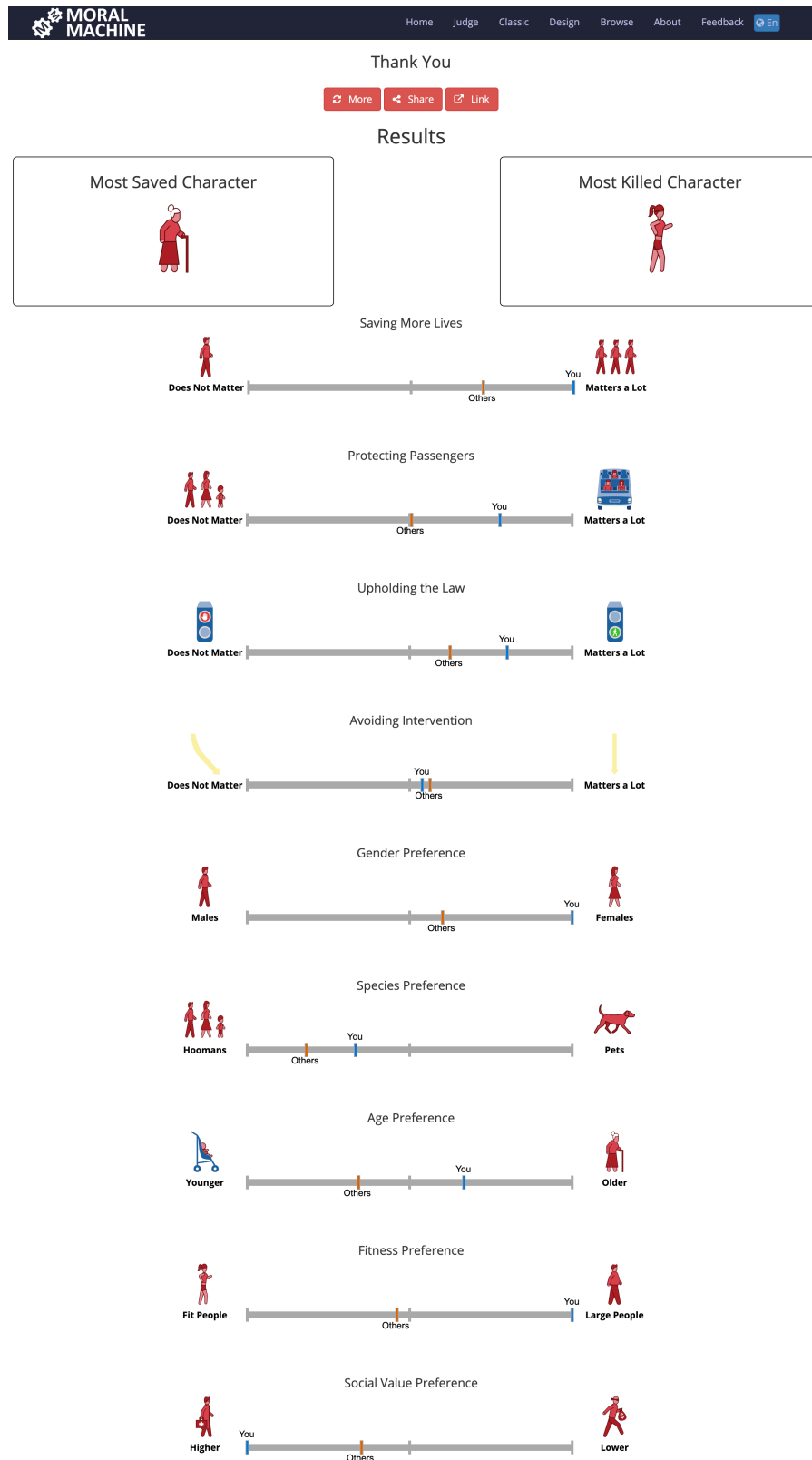
### III. Data & Code Availability

The data set from Moral Machine can be downloaded via: [https://osf.io/ukmzd/?view\\_only=4bb49492edee4a8eb1758552a362a2cf](https://osf.io/ukmzd/?view_only=4bb49492edee4a8eb1758552a362a2cf).


The entire project can be forked via: <https://github.com/eatxxl/Bachelor-Project-2019.git>.


Under this repository, there is a code folder that contains two python files: `moral_machine` and `correlation`. The former one for computing decision tree models, the latter one is for illustrating the correlation of data, both should be placed under a same file path along with the data which can be downloaded via the first link above.

## IV. Summary of Response



## V. Survey




[Home](#) [Jungle](#) [Classic](#) [Design](#) [Browse](#) [About](#) [Feedback](#) 

Would you like to help us better understand your judgement?


Yes

No


Disclaimer: These summaries are based on your judgment of a limited number of randomly generated scenarios, to help us keep the survey short. Therefore, these results are not definitive. Feel

Please answer the following questions. 


To what extent do you fear that machines will become out of control?

Very little  Very much


To what extent do you believe that your decisions on Moral Machine will be used to program self-driving cars?

Very little  Very much


To what extent do you feel you can trust machines in the future?

Very little  Very much


How willing are you to buy a self-driving car?

Very little  Very much

What are your political views?

Conservative  Progressive

What are your religious views?

Not Religious  Very Religious

Highest level of education


How old are you?

What is your gender?


Annual income, including tips, dividends, interest, etc (in US dollars)

From your point of view, how important was each factor in your judgement?  
(drag the sliders to where you think they should be)


Social Value Preference

Higher  Lower


Species Preference

Humans  Pets


Avoiding Intervention

Does Not Matter  Matters a Lot


Saving More Lives

Does Not Matter  Matters a Lot


Gender Preference

Males  Females


Fitness Preference

Fit People  Large People


Age Preference

Younger  Older

Protecting Passengers

Does Not Matter  Matters a Lot

Upholding the Law

Does Not Matter  Matters a Lot

Please describe any other rules here

Submit

Project by Scholastic Corporation or MTI Media Ltd