The Moral Machine - Computational Prediction of Human Choices

*Course: CSC 410*

*Researcher: Isabel Kathryn Ogilvie*

*Research Supervisor: Dr. Geoff Sutcliffe*

*University of Miami*

*Submitted Spring 2022*

1. **Introduction**

**1.1 The Moral Machine Experiment - Ethics of Artificial Intelligence**

Computational Morality, also known as machine morality, involves the analysis of decisions made by artificial intelligence. With the rise of artificial intelligence, technology’s influence on society and the decisions we make is becoming rapidly more prevalent, as we hand over the reins of our lives to computers. Accordingly, concerns about the ability of computers to make moral decisions in alignment with the ethics of human have come to the forefront. In response, many studies and research on the ethical principles of machines and the moral guidelines they should follow have cropped up. One such study is the Moral Machine Experiment, a thought experiment developed by a group at MIT exploring the moral problems encountered by self-driving vehicles.

Scenarios are presented to the user of their online platform in which participants must choose between saving the life of one group over another in a fatal car accident. The scenario’s structures are all variations on the trolley problem, where a self-driving car has two options: continue on the vehicle’s path and kill he passengers in the car or the pedestrians in the car’s path or alter its course and swerve to save the lives of those passengers but kill another group of pedestrians. The user then must choose whether they want to act to save the lives of the pedestrians/passengers who would be killed if the autonomous car chose not to act, or if they want to redirect the cars path to preserve the lives its transporting, at the expense of another group.

Scenarios are generated in groups of thirteen and contain at least two dilemmas focused on each of six dimensions of moral preferences: character species, age, gender, whether they are pregnant, their social status, their adherence to the law and their physical fitness. Although the scenario format remains the same, each scenario rotates through a various cast of characters, with a different combination in each scenario that tests the range of dimensions. Characters featured in the scenarios include individuals like girl, boy, Male Doctor, Female Athlete, Executive Female, Large Man, homeless, old man, old woman, dog, criminal, stroller, and so on.

**1.2 The Moral Machine Prediction Program**

This research project was very closely connected to the MIT Moral Machine, as originally it was a psych study into the difference ways STEM and Humanities students assigned worth to life. This was done by my fellow researcher, Anjou Sharma, a student in the University of Miami Psychology Department who used the Moral Machine simulation as a basis for her test, sending out a survey to University of Miami students in which they completed 20 of the aforementioned scenarios and then answered a series of follow up questions about certain factors’ influence on their decisions. There were enough computational components that she conducted the research under an additional supervisor outside of the Psychology Department, Dr. Sutcliffe of the Computer Science Department, who further expanded the project to include the aim of designing a computer model that could predict what choices individual might make in those scenarios based on the basic information they provided about their moral preferences.

If the prediction model succeeded, the intention was for designers of autonomous automotive systems to use these findings to ensure that vehicles are able to solve problems on the road that align with the moral values of humans around it.

For this purpose, the accuracy of the findings of the research were imperative. The basis behind the research was that if we were able to create a computational prediction model that could accurately guess what decision a person would make in a traffic accident scenario based on their preferences and values, we could equip the autonomous cars with this software and allow computers to make the decision exactly as a human would make it, but fast enough to act upon. Therefore, the main measure of success of this project would be the accuracy with which it could predict participant’s choices.

1. **Proposed Solution**
   1. **Comprehensive System Design**

Firstly, the survey described above was taken from a population of 99 University Miami students, in which they went through all 20 scenarios making their choices, and answered a comprehensive series of questions about how 7 moral preference dimensions influenced their choices (these dimensions will be discussed in detail later). The data from this survey would be the basis off which the prediction machine would be built, and act as a marker against which the accuracy of the predictions would be tested.

The data collected from the survey would then be divided into 2 encoded formats. The data from the first part of the survey, the scenario questions, were fed into a program that would convert the data from raw csv into a specially formatted csv format that could be easily understood by the prediction program as input. Note that the encoded scenarios only contained the characters, not the user’s choices on who to save – the survey participants’’ choices would be stored somewhere else and then checked against the choices generated by the program for each user later. The data from the second part of the survey, the detailed questionnaire in which the participant assigned importance and priority to save to the moral dimensions and values respectively, was fed into a separate program to convert it from raw csv to a more readable csv format which would be then input to the prediction model.

Once the data was encoded, it would be fed into the Java prediction program, which would predict which group each user would choose to save in the encoded scenarios based on their moral preferences. A mathematical model designed by Dr. Sutcliffe was the basis for calculating the predictions and will be explained in more detail in the next section. Once the program generated the predictions, the results would be cross checked against the actual choices made by participants to verify the accuracy of the program. A detailed diagram showing the system components is shown below in Figure 1.

Diagram

Description automatically generated

*Figure 1: Moral Machine System Components Diagram*

It was also planned to increase the accuracy of the program by using a larger data asset than the 99 University students by using web scraping to gather further participant data from the Moral Machine Experiment website, moralmachine.net and scraping all the responses submitted to them. All the collected responses would then be fed into the encoders and input to the moral machine, which in theory would be able to arrive at a more accurate analysis with a larger data set.

**3. Research Algorithm Design**

**3.1 Dimension and Value / Scenario Data Encoder Creation**

There were two encoders to be written: the Dimension and Value (D&V) encoder, and the Scenario Encoder. The function of these was to parse only the necessary fields of data from the wide array of fields collected in the original survey. The necessary fields from the D&V encoder were: Participant ID, demographics, all dimension related data, and all values related data. An example of the correct encoded format is as shown in Figure 2. The second encoder was the Scenario Encoder, the function of which was to encode data on the participants on each group of the scenario and their description (namely the 7 dimensions they belonged to). An example of an encoded scenario can be seen in Figure 3. Although the encoders were initially intended to be written in Java, they were ultimately coded in R instead due to technical difficulties during the project execution.

Text

Description automatically generated

*Figure 2: Dimension and Value Encoded Format Example*

Text

Description automatically generated

*Figure 3: Scenario Encoded Format Example*

**3.2 Moral Machine Mathematical Model**

The Moral Machine Prediction Program was based on a simple mathematical model, which was to be refined or redesigned if the results proved necessary. The data collected from the users was split into 2 sections: the user’s moral preferences and the scenarios’ members. For the survey participant’s preferences, they were asked to assign a value of importance to 7 factors in their choices to save or kill. The first was Type, in other words whether the scenario member was a human or animal. Next was Age, with Gender as the third type. They were also further asked to rank the importance of a character’s pregnancy status (i.e., whether they were pregnant or not). Lawfulness was the fifth factor, which described whether the character was breaking any traffic laws or not. The sixth dimension was Social Status, simply their rank or ascribed societal value; for example, a doctor or executive would be a professional, whereas a criminal or someone homeless would be a failure. Lastly, the seventh dimension was health, or in other words, the character’s physical fitness level. Each dimension had subcategories to it, which would each be later ascribed a “value”, as seen below in the list of dimensions and their corresponding values:

* Type - Human, Animal
* Age - Baby, Child, Adult, Elderly
* Gender - Male, Female
* Pregnancy - Pregnant, Not Pregnant
* Lawfulness - Passenger, Abiding, Breaking
* Social Status - Professional, Average, Failure
* Health - Athlete, Average, Obese

Participants could assign each category as Important, Somewhat Important or Not Important, with corresponding numerical values of 2, 1 and 0 respectively in the mathematical model based on how much they felt that specific category influenced their decision. After assigning an overall numerical value to each main category (Type, Age etc.), they were then asked to rank each subcategory based on who they felt they should save the most in the following order: Willing to Kill, Indifferent, Kinda Want to Save and Must Save, with corresponding numerical values of -1, 0, 1, 2 in the mathematical model. Each scenario had two groups, swerve and straight, each of which had one or more lives in it which would have value for all these categories. The program would take these values for each life and then compute the individual’s overall numerical worth, or life saving value, (LSV) by the following formula:

For dimension switch an unknown value, the average of the known values was assigned (for example, if the sex VV for male is M and sex VV for Female if F, a baby’s gender would be assigned the value of (M + F) / 20. In the case of an inapplicable value, a numerical value of 0 was assigned. For example, the social status VV for a baby would be set to 0. An example scenario and according calculation is as follows:

*1. Type is important, DMF\_type = 2*

*Really want to save Humans, VV\_human = 2*

*Indifferent about animals, VV\_animal = 0*

*2. Age is somewhat important, DMF\_age = 1*

*Really want to save adults, VV\_adult = 2*

*Kinda want to save children, VV\_child = 1*

*Indifferent about babies, VV\_baby = 0*

*Willing to kill elderly, VV\_elderly = -1*

*3. Gender is not important, DMF\_gender = 0*

*Really want to save females, VV\_female = 2*

*Kinda want to save males, VV\_male = 1*

*4. Pregnancy is somewhat important, DMF\_pregnancy = 1*

*Really want to save non-pregnant, VV\_non-pregnant = 2*

*Kinda want to save pregnant, VV\_pregnant = 1*

*5. Social status is somewhat important, DMF\_social-status = 1*

*Really want to save average, VV\_average-ss = 2*

*Kinda want to save professional, VV\_professional = 1*

*Willing to kill failures, VV\_failure = -1*

*6. Lawfulness is important, DMF\_lawfulness = 2*

*Really want to save law-abiding, VV\_law-abiding = 2*

*Indifferent about law breaking, VV\_law-breaking = 0*

*7. Health is somewhat important, DMF\_health = 1*

*Really want to save average, VV\_average-health = 2*

*Kinda want to save athlete, VV\_athlete = 1*

*Indifferent about obese, VV\_obese = 0*

*Group contains:*

*1. Law-abiding male adult:*

*LSV = (2\*2 + 1\*2 + 0\*1 + 1\*NA + 1\*2 + 2\*2 + 1\*1) / 6 = 2.17*

*2. Baby:*

*LSV = (2\*2 + 1\*0 + 0\*1.5 + 1\*2 + 1\*NA + 2\*NA + 1\*1) / 5 = 1.00*

*3. Pregnant female adult executive law-abiding athlete:*

*LSV = (2\*2 + 1\*2 + 0\*2 + 1\*2 + 1\*1 + 2\*2 + 1\*1) / 7 = 2.0*

**3.3 Moral Machine Prediction Program Design**

The Moral Machine prediction program was designed as an object-oriented system, composed of 2 main objects each containing several smaller objects. The first main object was Preference, which described the participant’s moral preferences. It contained the numerical values of overall importance they assigned to each of the 7 dimensions (Type, Age, Gender, Pregnancy, Lawfulness, Social Status, and Health), as well as the ranking for each of the subcategories for the dimensions (the values).

**4.** **Moral Machine Prediction Program Prototype**

**5. Choices of Technology**

We chose to use the java coding language as it was the language the coder was most familiar with. The project was largely object oriented, being centered around the two main objects, Preferences and Group, with the two smaller objects people. Therefore, we needed to choose an object-oriented programming language. Java is a programming language with a robust object-oriented programming capability. IntelliJ was the Java Programming environment chosen.

We also used R-Studio for the collection and organization of the survey data collected from all the participants. The survey data from 100 students was collected and stored in R in the form of excel sheets, from which it was stored in and organized into csv files of the correct encoder format as mentioned earlier. Although originally it was intended to carry out the encoder function in java, we switched over to R due to some technical difficulties, as well as R’s innate data organization capabilities. The R programming language is designed for statistical computing, and is tailor made for data science projects much like this one, and as thus it is appropriate for organizing large data sets and the analyses of said data sets. From there it was easy to transfer the encoded material into R. <Used the C language>

**9. Conclusion**

 For this purpose, the accuracy of the findings of the research were imperative. The basis behind the research was that if we were able to create a computational prediction model that could accurately guess what decision a person would make in a traffic accident scenario based on their preferences and values, we could equip teh autonomous cars with this software and alow computers to make the decision exactly as a huan would make it, but fast enough to act upon.

 persons go through fatal car accident scenarios on an online platform where they are made to choose between saving one group of people and condemning the other to death. The autonomous vehicle in