THE MORAL MACHINE EXPERIMENT: PREDICTING MORAL DECISION-MAKING BASED ON PERSONAL VALUES

A COMPARISON OF RANDOM FOREST, SUPPORT VECTOR MACHINES AND K-NEAREST NEIGHBORS

TERESA VIRCA

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2090933

COMMITTEE

dr. Michal Klincewicz
dr. Mojtaba Rostami Kandroodi

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

8.795

THE MORAL MACHINE EXPERIMENT: PREDICTING MORAL DECISION-MAKING BASED ON PERSONAL VALUES

A COMPARISON OF RANDOM FOREST, SUPPORT VECTOR MACHINES AND K-NEAREST NEIGHBORS

TERESA VIRCA

Abstract

This research employs the dataset of the Moral Machine Experiment by Awad et al. (2018) to investigate the influence of personal values on decisions within the context of autonomous vehicle moral dilemmas. Unlike previous works, which primarily focus on culturally homogeneous groups, this study examines responses within a culturally diverse sample. This makes it possible to isolate personal values from cultural values when predicting moral decisions, as this separation is fundamental according to value scholars (Hofstede, 1984; Schwartz, 2006). Building on the computational model for extracting abstract features devised by Kim et al. (2018), this research investigates the predictive performance of Random Forest, Support Vector Machine, and K-Nearest Neighbors, alongside a dummy classifier, as their potential in relation to the MME remains understudied. The findings suggest that all three models are able to identify patterns in the relationship between personal values and moral decisions, in particular considering factors such as species, legality and age. Moreover, predictive performance improves with increased sample size and in culturally homogeneous samples. This last finding suggests that when studying culturally homogeneous groups, predictions should be understood as the product of both personal and cultural values. This thesis contributes to existing research on values and moral decision-making, as well as to the data science domain.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

- **Data Source**: The data utilized in this research originates from the Moral Machine Experiment, detailed in the publication by Awad et al. (2018). Access to the data is publicly available at https://goo.

gl/JXRrBP. Since the project makes use of open data, making this thesis publicly available is in line with the reasoning of the original authors. The dataset has been anonymized, ensuring the privacy of respondents. Moreover, the demographic data of the respondents has not been included in the analysis. As this study solely relies on pre-existing data, no direct involvement with human participants or animals occurred during its execution.

- **Figures**: All figures utilized in this thesis have been created by the author, except for Figure 1, Figure 2, and Figure 6. These consist of screenshots from an online web-page, and they are appropriately cited according to APA citation style.
- **Code**: The code employed for data analysis and model development has been independently developed and is publicly available on GitHub (https://github.com/Moral-Machine-2024/t.virca.git). This codebase can be used to replicate the findings of the thesis. The program and libraries used to generate the outputs are provided in the methodology section below.
- **Technology Tools**: The thesis was formatted using Overleaf, and Zotero was used for reference management throughout the research process.

2 PROBLEM STATEMENT AND RESEARCH GOALS

2.1 *Context*

The present research relies on data gathered for the Moral Machine Experiment (MME) developed by Awad et al. (2018). The experiment consists of moral dilemma scenarios which resemble the structure of trolley problems first introduced by Philippa Foot (1967) and later used by Judith Jarvis Thomson (1976). The experiment, which went viral online and gathered 40 million responses from all continents, presents users with a scenario where an autonomous vehicle (AV) must decide whether to swerve or stay on course, thereby sacrificing and sparing different groups of individuals. Given the great amount of information provided by the MME, this research delves into popular understandings of machine morality and employs the openly available data to extract further insights in the field of ethical decision-making involving machines.

2.2 *Motivation*

Today's society faces the proliferation of AI systems. Machines are increasingly deployed to perform tasks previously reserved for humans, and are therefore designed to imitate human behavior and make decisions on their behalf (Seifert et al., 2022). Since these decisions sometimes have ethical implications, AI ethicists aim to understand how machines can behave morally, and which considerations should drive their development (Berberich & Diepold, 2018; Charisi et al., 2017; Moor, 2006). Additional considerations relate to the value alignment problem first described by Wiener (1960). This refers to the challenge of ensuring that autonomous machines are guided by human values however unreliable, nuanced and implicit these may be. Consequently, a growing body of literature addresses the technical, legal and ethical ramifications of the problem, while attempting to formulate solutions (Arnold et al., 2017; Christian, 2021; Hadfield-Menell & Hadfield, 2019; Peterson, 2019; Sierra et al., 2021). Since autonomous systems are intentionally or unintentionally imbued with human values, and since these are sometimes destined to make ethical decisions, it is important to understand the relationship between individual values and ethical decision-making (Wallach & Allen, 2008). To this end, the data provided by the MME is particularly useful, since it provides a large collection of value preferences and related moral decisions. Given its quantity, the data is especially fit for capturing the nuisances of human moral judgement and values, as Bourgin et al. (2019) note that human behavior is noisy and diverse, and it therefore requires large sample sizes to be appropriately captured.

Moreover, the scenarios provided by the MME are potentially realistic ethical conundrums involving self-driving cars, as it is sometimes unfeasible to simultaneously maximize the safety of passengers, surrounding motorists and pedestrians. Goodall (2014) suggests that since AVs are undoubtedly going to find themselves in unavoidable crush situations, and since the moments preceding the crush require the vehicle to make moral decisions, ethical algorithms are bound to develop as technological capabilities advance. Given this inevitability, it is important to study the influence of values which are used to develop benchmarks for machines' ethical decision-making, as the ones created by Kim et al. (2018). Besides the societal relevance of this topic, this research aims to contribute to the data science domain, as it employs computational modeling to extract personal values and machine learning to analyze their predictive capabilities. Even though the state-of-the-art accuracy is not surpassed, this research employs a novel conceptualization, methodology, and provides a direction for future research.

2.3 *Problem statement*

The data of the MME provides a unique opportunity to combine the literature on human values with machine learning. Despite this, existing studies using the data, do not explicitly focus on values. In the original experiment, Awad et al. (2018) conduct an exploratory analysis of the data and present findings on moral preferences across various clusters of countries. They find that conceptions of morality differ across cultures. In fact, the authors draw parallels with the dimensions of cultural values created by Inglehart and Welzel (2005), which include for example, individualistic and collectivistic cultures. Conclusions such as *"participants from collectivistic cultures [...] have a weaker preference for sparing younger characters"* are undoubtedly compelling (Awad et al., 2018, p. 8). However, scholars studying values caution against making generalizations about individual responses based on culture, since personal and cultural values lie on two distinct dimensions (Hofstede, 1984; Schwartz, 2006). Indeed, considering otherwise produces an ecological fallacy (Robinson, 1950). Existing studies make precisely this mistake when devising models which identify patterns in culturally homogeneous samples in order to predict individual decisions (Kim et al., 2018; Wiedeman et al., 2020). Given these insights, this thesis studies moral preferences in the MME in relation to existing literature on values. Particularly, it aims to disentangle personal values from cultural values by creating a predictive model focusing on culturally diverse samples of respondents. The conceptualization of personal values derived from the variables of the MME is further explained as part of the literature review.

Only a few papers use machine learning to study the MME. The authors of the original experiment develop a hierarchical Bayesian model which enables them to predict moral decisions based on "moral principles" (Awad et al., 2018; Kim et al., 2018). Wiedeman et al. (2020) propose an improvement to the Bayesian approach which introduces non-linearity by means of deep learning, thus producing the state-of-the-art accuracy (0.80). Finally, Agrawal et al. (2019) employ tailor-made models to produce more transparent predictions, that can be more easily tied to human moral reasoning. The limited range of models used on the data highlights a gap in the literature. Therefore, the present research examines the performance of different machine learning algorithms for predicting moral decisions based on personal values. Studies in behavioral science focused on predicting humans decisions often compare the effectiveness of Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) (Bourgin et al., 2019; Plonsky et al., 2017, 2019). Hence, their performance is examined against that of a naive classier.

2.4 *Research question and research strategy*

To address the gaps in the literature, this research proposes the following research question:

> *What influence do personal values have on the decision to save or sacrifice characters in the context of dilemma situations where an autonomous vehicle has to decide whether to swerve or stay on track?*

The following sub-questions help answer the main question:

RQ1 *Are the following personal values predictive of the decision to save or sacrifice characters in culturally diverse samples: (1) intervention, (2) gender, (3) age, (4) pregnancy, (5) fitness, (6) social status, (7) species, (8) the distribution of harm between passengers and pedestrians, and (9) the legality of actions?*

RQ2 *Which of the following models is able to best predict decisions based on personal values: Random Forest, Support Vector Machine or K-Nearest Neighbors?*

RQ3 *Does the predictive accuracy of personal values increase with increased sample size?*

RQ4 *Is the predictive performance of values higher in culturally homogeneous samples?*

RQ5 *Which of the listed personal values are more important for predicting decisions?*

This paper investigates the influence of personal values on moral decisions in a culturally diverse sample, by testing the effectiveness of RF, SVM and KNN in comparison to a dummy classifier. The sample chosen is culturally diverse, as it includes responses from all continents and from all cultures, as categorized by the Inglehart-Welzel Cultural Map 2010-2014 (Awad et al., 2018; Inglehart & Welzel, 2005). Employing a culturally diverse sample serves to limit the models' dependency on patterns derived from cultural values, and instead solely rely on individuals' personal values. A dummy classifier is used as baseline, and the evaluation metrics of three additional models with diverse algorithmic characteristics are compared. The first sub-question lists the personal values considered as predictors. The second sub-question specifies the models to be considered. The third sub-question investigates whether increased sample size leads to increased classification accuracy, as a way to ascertain the predictive capability of the data. The fourth sub-question attempts to verify the assumption made about cultural homogeneity and heterogeneity. That is, whether predictive performance

improves in light of additional patterns derived from similarities in cultural values. To this end, the best performing model is trained and used to make predictions on three country samples of similar size. Finally, the fifth sub-question aims to understand the relative influence of personal values, as permutation feature importance is performed to shed light on the most and least predictive features.

2.5 *Findings*

The results suggest that increasing sample size up to 10^5 observations improves the models' predictive performance, in line with expectations (Agrawal et al., 2019; Kim et al., 2018). Even though the accuracies obtained here do not surpass benchmark accuracies in the literature, all three models are able to learn patterns based on personal values, and the SVM algorithm produces the highest accuracy score. The values *species*, *legality*, and *age* are especially used by the model to predict decisions, however, this result is influenced by the frequency with which these features appear in the scenarios. Lastly, the best SVM model displays an increased predictive performance in the three culturally homogeneous samples tested, thus indicating that additional patterns are identified based on cultural similarities. Overall, this research concludes that individuals' personal value hierarchies influence their moral judgement, and can be thus used in prediction tasks. Additionally, when studying culturally homogeneous groups, predictions should be understood as the product of both personal and cultural values.

3 LITERATURE REVIEW

The literature suggests that important determinants of moral decision-making include individuals' moral reasoning and values (Blasi, 1980; Ravlin & Meglino, 1987; Rest, 1986; Rokeach, 1973; Weber, 1993). Despite this, existing studies employing the data of the MME do not make use of these determinants as part of their conceptual framework. Therefore, the following sections review the literature on moral reasoning and values, and apply them to the MME. Additionally, the relationship between values and culture is highlighted, as it represents a central component of this thesis. Finally, a section is reserved for the introduction of computational models developed for the MME, alongside their respective accuracies.

3.1 *Trolley dilemmas and moral reasoning*

Trolley dilemmas are often used by scholars to isolate moral judgements. The literature suggests that variations in situational and psychological factors, determine what is viewed as the most moral of two alternatives (Baron & Ritov, 2004; Greene et al., 2009; Royzman & Baron, 2002). For instance, in the original trolley problem it is considered permissible to push a lever and sacrifice the life of one man in order to save those of five men (Foot, 1967). On the other hand, in Thomson's variation of the dilemma, it is less often considered permissible to push a large man off a bridge in order to stop the trolley (Thomson, 1976). The moral judgements arising from the two scenarios might seem incongruent from a consequentialist perspective, since the final outcome is the same, however, comparing the two highlights that contextual elements within the scenario can alter respondents' perception of what is morally acceptable. This remains true for the scenarios in the MME, whose realistic context allows respondents to factor-in personal and deontological perspectives alongside the more commonly invoked utilitarian one. That is, considering not only the consequences of an action, but also categorical maxims such as *"one ought to follow the rules of the road"*.

Even though findings by Bostyn et al. (2018) suggest that responses to dilemma scenarios are not representative of behavior in real-life situations, Plunkett and Greene (2019) emphasize that dilemma scenarios need not emulate reality, since they primarily serve as a means to grasp the cognitive processes behind moral reasoning. In fact, the scenarios of the MME, which require respondents to ponder on the right course of action, are more representative of what moral psychologists define as type II cognition (Luft, 2020). This type of cognitive processing relies on explicit and discursive deliberation as opposed to intuition and gut feeling (Haidt, 2001). Blasi (1980) defines moral reasoning as the human capacity to *"wonder about the fundamental criteria for right and wrong, compare different criteria, and inquire about their correctness and truth value"* (p. 3). The contribution by Luft (2020), is important for highlighting the role of context and culture in shaping moral reasoning. She notes that this wondering about right and wrong is determined not only by the action itself, but also by the environment where the action takes place, and by the social nature of the interaction. Therefore, the data provided by the MME is important not only for extracting moral reasoning, but also for relating this to cultural and contextual variables. Because of the prerogatives of trolley-like dilemma scenarios, this research considers moral decisions within the MME as the result of cognitive processes which shape moral reasoning and are inextricably related to personal values (Weber, 1993).

3.2 *From moral reasoning to values and culture*

The literature studying the influence of moral reasoning on judgement is supplemented by further studies identifying the influence of values on reasoning (Rest, 1986), the influence of reasoning on values (Frederick, 2005; Pennycook et al., 2014), and the influence of values alongside reasoning (Blasi, 1980; Ostini & Ellerman, 1997; Rokeach, 1973; Weber, 1993), clearly highlighting the interdependent nature of values and moral reasoning. Studies which focus on moral reasoning alone, still emphasize the importance of personal values in helping individuals to navigate moral landscapes and align with societal norms (Rest, 1986; Weber, 1993). Analyzing the relationship in the opposite direction, "reflectionists" show that individuals' moral reasoning shapes the subject-matter of their held beliefs (Frederick, 2005; Pennycook et al., 2014). In particular, Weber (1993) finds that *"the individual's moral reasoning process is the vehicle used to activate, filter and translate personal values into behavior"* (p. 457). In order to study moral judgement it is therefore important to consider existing work on personal values. In his seminal work, Rokeach (1973) defines values as *"an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence"*. This often cited definition points to several important attributes, including (1) the distinction between instrumental and terminal values, denoted by *"mode of conduct"* and *"end-state of existence"*, (2) the formation of a hierarchy of values within the individual, and (3) the enduring state of the belief, which distinguishes values from more volatile personal traits (Ostini & Ellerman, 1997; Sagiv & Schwartz, 2022; Schwartz, 2006; Weber, 1993).

Feather (1988) further highlights that individuals' cognitive processes and value hierarchies are determined by past socialization experiences, which are necessarily linked to cultural factors. In fact, a portion of the literature on this topic focuses on mapping value emphases across different societies and considers them the central component of culture (Hofstede, 1984; Inglehart, 2020; Schwartz, 2006). Even though the link between personal and cultural values is evident, these do not necessarily overlap within an individual. Hofstede (1984) and Schwartz (2006) view personal and cultural values as belonging to fundamentally distinct dimensions, since considering otherwise would lead to an ecological fallacy, as explained by Robinson (1950). Considering a practical example, Elster and Gelfand (2021) show that the relationship between cultural values and behavior is stronger in societies characterized by stricter norms of conduct. Conversely, behavior is not as affected by cultural values in societies characterized by relaxed norms, thus firmly disconnecting personal and cultural values (Sa-

giv & Schwartz, 2022). Given the relevance of values and culture for moral decision-making, this thesis exploits the vast collection of value preferences provided by the MME to study the influence of personal values on moral judgement in a culturally diverse sample. More specifically, focusing on a diverse sample ensures that the patterns relied on for predicting decisions are primarily representative of personal rather than cultural beliefs.

3.3 *Personal values in the Moral Machine Experiment*

The MME by Awad et al. (2018) was designed to capture respondents' considerations to moral questions by including numerous variables, which when pitted against each other, are able to highlight inherent preferences. These variables span three "structural factors" and twenty types of characters with various demographic attributes. Awad et al. (2018) explain that "structural factors" describe the environment of the scenario, including whether the vehicle should swerve or stay on track, whether pedestrians are pitted against passengers in the vehicle or against pedestrians on the other side of the street, and whether pedestrians are crossing the street on a green or a red traffic light, that is, legally or illegally. The characters are assigned attributes including gender, age, pregnancy, fitness, social status and species (Figure 1 provides a visualization of the characters in the experiment). When faced with a scenario a user could decide to always save the greatest number of individuals and always prioritize the lives of certain characters. However, a scenario might require the user to abandon one of the two preferences, when both options violate initially considered predilections. By requiring respondents to establish a trade-off between preferences, a hierarchy is formed. This outcome of the experiment resonates with the notion of value hierarchy, according to which values that are deemed more important are more likely to steer decision-making (Blasi, 1980; Ravlin & Meglino, 1987; Rokeach, 1973; Sagiv & Schwartz, 2022). An example of a scenario is displayed in Figure 2.

Decisions in this context follow a utilitarian logic where a trade-off between different values has to be made in order to identify the outcome with the maximum utility. In order to extract the values inherent to each scenario, the variables described above are deconstructed into several abstract components, which can better represent respondents' beliefs. Drawing inspiration from the theoretical model devised by Kim et al. (2018) the abstract components obtained include *intervene*, *male*, *female*, *young*, *old*, *infancy*, *pregnancy*, *fat*, *fit*, *working*, *medical*, *homelessness*, *criminality*, *human*, *non-human*, *passenger*, *law abiding* and *law violating* (see Figure 3). This deconstruction makes it possible to assign the same abstract component to multiple characters, and vice versa, assign the same character to multiple

Figure 1: Twenty character icons

Figure 1: From *Moral Machine*, n.d. by E. Awad, S. Dsouza, P. Chang. (www.moralmachine.net)[screenshot].

abstract components. For example, if value is placed on females, many characters including working women, medical professionals, pregnant women, athletes and little girls are to be considered. Vice versa, if value is placed on female doctors, both the concept of female and that of doctor are encompassed. In this context, personal values are understood as the relative importance given to each abstraction by the respondent. For instance, placing value on the *medical* abstraction translates into the belief that doctors are essential for the functioning and prosperity of a society, and that their lives should thus be prioritized. When considering the definition provided by Rokeach (1973), such belief represents an instrumental value, which is conducive to the well-being of society. Furthermore, the belief is considered to be "enduring", as it is unlikely to continuously change based on context and time, as is the case for traits, needs, and motives (Sagiv & Schwartz, 2022).

Research in moral and cognitive psychology supports the importance of abstractions for the ordering of values. Blasi (1980) suggests that the interaction of superordinate concepts is a fundamental component of morality. This is because when relying on rational processes, much like with type II cognition, meaning is constructed based on value hierarchies with reference to abstractions (Blasi, 1980). In the same vein, Kohlberg

Figure 2: Example scenario

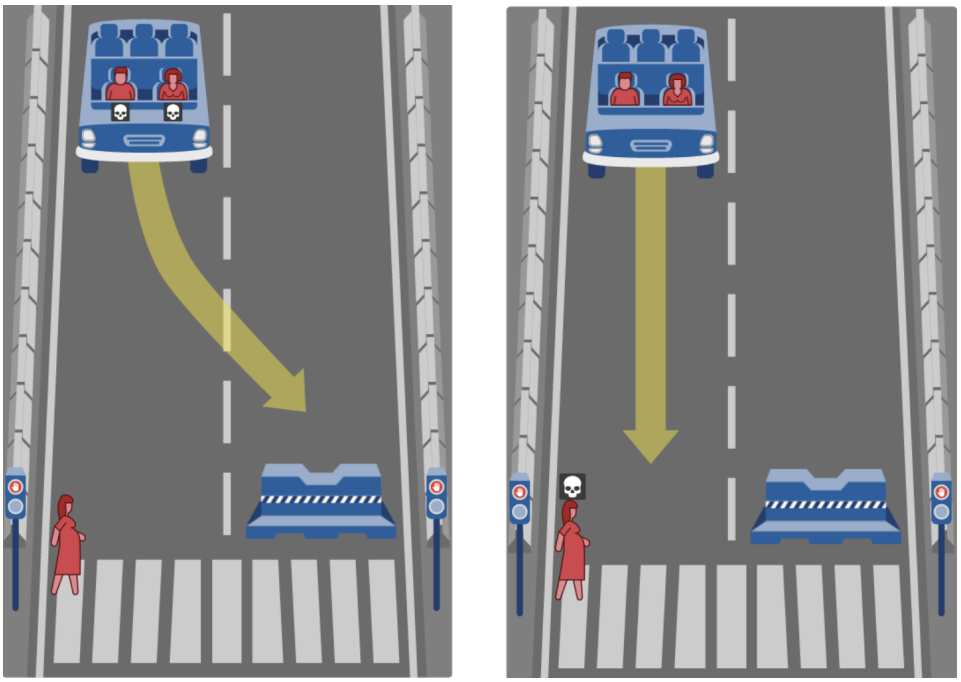


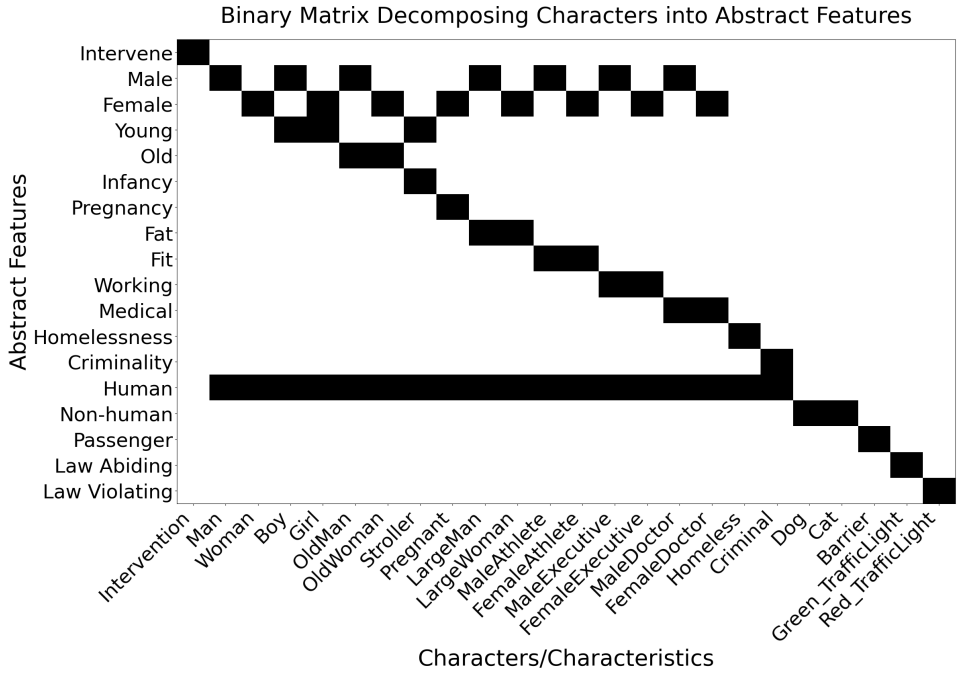
Figure 2: From *Moral Machine*, n.d. by E. Awad, S. Dsouza, P. Chang. (www.moralmachine.net)[screenshot].

(1981) finds that children rely on abstract concepts when facing moral problems, as a mechanism which helps them make sense of the world. Additionally, studies focusing on computational cognitive modeling often rely on these or similar assumptions to create mathematical models to imitate or better understand human cognition (for example see Agrawal et al., 2019; Mikhail, 2002, 2007; Sun, 2009). Building on these considerations, the methodology section will further illustrate the feature extraction process transforming variables into abstract features representative of personal values.

3.4 *Computational modeling and model comparison*

Only a few studies have employed MME data in combination with computational modeling. Relevant contributions are provided by Kim et al. (2018), Noothigattu et al. (2018), Wiedeman et al. (2020), and Agrawal et al. (2019), as their models are aimed at predicting respondents' decisions based on the variables provided. The methodology employed by Kim et al. (2018) is particularly important for this research, as their theoretical model for extracting abstract features was replicated in this thesis, albeit with a

Figure 3: Deconstruction of variables into abstractions



unique practical implementation. They develop a hierarchical Bayesian model which allows them to accurately infer decisions based on vectors of what they define as "moral principles". The study models various benchmarks aimed at predicting the moral judgement of single individuals and culturally homogeneous groups, that is, samples of respondents from the same country. They obtain benchmark accuracies ranging from 0.70 to 0.75 for decisions made by a single individual. Noothigattu et al. (2018), build on the work by Kim et al. (2018) to create a theoretical model to systematize ethical AVs decision-making by combining methods derived from machine learning and computational social choice. In particular they perform pairwise comparisons for respondents and aggregate the individual models to extrapolate collective preferences. Their methodology is inspired by Conitzer et al. (2017), who discuss game theory and machine learning approaches for formalizing a general theoretical framework for AI moral decision-making. Wiedeman et al. (2020) uses the same methodology presented by Kim et al. (2018) which relies on vectors of "moral principles", and applies it to an artificial neural network composed of various dense layers. The network is able to obtain accuracies ranging from 0.70 to 0.80. Finally, Agrawal et al. (2019) make use of the data to develop a transparent methodology for better understanding the relationship between the model's predictions and human reasoning. To this end, they construct

models by gradually adding new types of features so as to control their interpretability, and then compare their performance to that of a neural network with 32 hidden layers. Their models obtain accuracies ranging from 0.57 to 0.77 when using a neural network.

Taking inspiration from existing studies, this research compares the predictive performance of three models with different algorithmic characteristics against a naive classifier which does not take into account the class distribution in the training data. The three models are popular in machine learning, and they are often used in the field of behavioral and cognitive science to model human behavior and decision-making. For example, Plonsky et al. (2017, 2019) and Bourgin et al. (2019), compare the performance of RF, SVM and KNN when predicting human choices. The comparison of the three is interesting as they represent three distinct learning methods. RF is a type of bagging ensemble method, where the prediction of multiple decision trees is aggregated (Breiman, 2001; Dietterich, 2000). SVM is a kernel method which identifies the hyperplane that separates the outcome classes in the feature space as optimally as possible (Burges, 1998). KNN, on the other hand, is a non-parametric and instance-based model which makes predictions by selecting the majority class among the nearest neighbors (Cover & Hart, 1967). Regarding the suitability of the models, while employing different learning methods, all three are able to handle categorical features and to capture non-linear relationships in the data (Agresti, 2012).

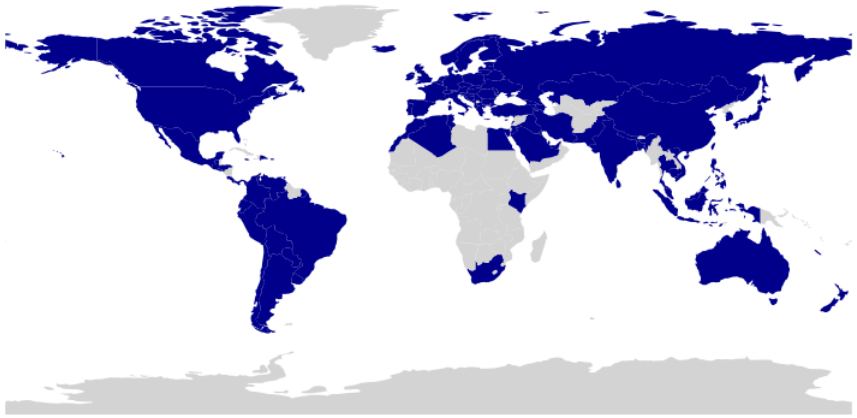
4 METHODOLOGY AND EXPERIMENTAL SETUP

4.1 *Sampling and dataset description*

The original dataset of the MME contains 39.61 million decisions from respondents in 233 countries (Awad et al., 2018), amounting to a file of 11.6 GB. Given that the research question here analyzed requires a culturally diverse pool of respondents, sampling is conducted by extracting countries which are assigned a culture according to the Inglehart-Welzel Cultural Map used by Awad et al. (2018). Cultures include *Orthodox*, *Islamic*, *Latin America*, *English*, *Catholic*, *Protestant*, *Confucian*, *Baltic* and *South Asia*. The cultural map is based on information gathered from the World Values Survey, a global research project that explores values and beliefs across various dimensions from countries around the world (Inglehart & Welzel, 2005; WVSA, 2020). The selected countries can be viewed in Figure 4. Importantly, a separate dataset contains the culture information, which can be retrieved within the same repository of the MME. As the outcome of the extraction and merging procedure, six samples are created corresponding

to each continent - Africa, Asia, Europe, North America, South America and Oceania - to which the culture information is added. The six samples are of varying sizes, in particular Europe and North America contain significantly more observations than Africa and Oceania.

Figure 4: Countries included in the final sample



As the size of the six samples remains substantial, data cleaning steps are performed separately on each before they are merged. Steps include deleting unnecessary columns so as to reduce computational complexity, renaming columns for clarity, and deleting rows containing missing values in any variable to be employed in future engineering steps. The amount of missing values remains neglectable relatively to the overall size of each sample. Importantly, each scenario contains two outcomes, that is the outcome where the vehicle stays on course and the outcome where the vehicle swerves. Each outcome is encoded in a separate row of the dataset, detailing the presence of "structural" and character features, as well as the decision of the respondent to save or sacrifice the characters in that outcome. The two rows can be paired by means of a common response ID, which makes it possible to obtain a complete scenario. Given this structure, an additional data cleaning step filters only those outcomes with a matching response ID, thus filtering out incomplete scenarios.

For the purpose of size reduction, two functions were defined, one for sampling based on given criteria (*create_sub_sample*) and one for verifying the distribution of values after sampling (*verify_proportions*). The former operates by taking a dataset as input and generating a sub-sample that accounts for a specified proportion of the original sample size. This is achieved through the implementation of stratified shuffle-split, ensuring that the sampling process is stratified based on country, which equally

ensures that the stratification applies to culture. Furthermore, the function is designed to include both sides of a scenario when selecting the indices of the chosen rows, maintaining the integrity of the dataset's structure. The latter function verifies the correct functioning of the former, by displaying the count and proportion of countries and cultures in the original sample and the created sub-sample. Additionally, it displays the distribution of the target variable *saved*, to ensure that both sides of a scenario are selected. This is because, within a scenario, an outcome is necessarily chosen and another discarded by the user. These functions are systematically applied to each of the six samples to initially halve their sizes. Table 1 displays the outcome of the second function verifying the proportion of the culture variable in the *Asia* dataset (D1) and outcome sub-sample (D2).

Table 1: Example function outcome for Asia dataset

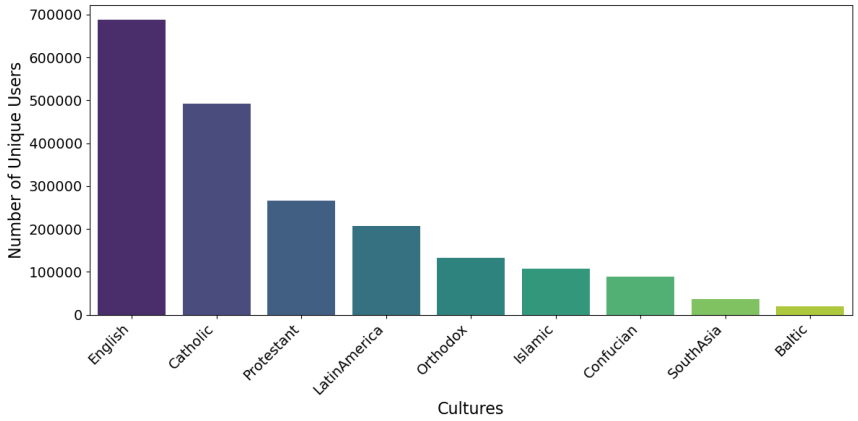
Culture	**Count D1**	**Proportion D1**	**Count D2**	**Proportion D2**
Islamic	2.545.130	0.3947	1.272.564	0.3947
Confucian	2.511.632	0.3895	1.255.818	0.3895
SouthAsia	1.105.112	0.1714	552556	0.1714
LatinAmerica	227.956	0.0353	113.978	0.0353
Orthodox	57.258	0.0088	28.628	0.0088

Finally, this step allows for the merging of the sub-samples into one dataset containing responses to complete scenarios from all continents of the world, and maintaining a proportional distribution for the culture variable. Figure 5 shows the distribution of unique users across the various cultures within the merged sample, which contains over 16 million scenarios. This dataset alongside the above mentioned functions are used to further extract country samples and samples of increasing size to be used during the analysis.

4.2 *Data Pre-processing: Feature Engineering*

The decision of the respondent to save or sacrifice the characters in a given scenario is represented by variable *saved*, characterized by two possible values $\{0, 1\}$, where $saved = 1$ saves the characters in the outcome, whereas $saved = 0$ sacrifices the characters. In Figure 6 the vector $\boldsymbol{\theta}_1$ encodes the characters and "structural variables". The example outcome displayed contains a one for *intervention*, one *man*, one *old man*, two *old women*, and a *red traffic light*. So the presence of "structural variables" is represented by a

Figure 5: Distribution of cultures

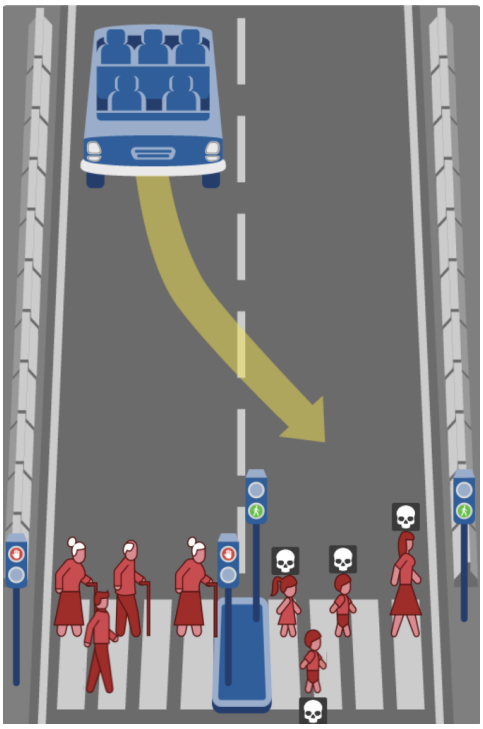


one as opposed to a zero, and the presence of the characters is represented by the respective count.

The vector $\boldsymbol{\theta}$ is translated into the vector $\boldsymbol{\lambda}$ by means of a linear transformation, forming a binary matrix of size 18×24 defining the latent feature space, as displayed in Figure 3. The length of the latter vector is reduced, as each original variable belongs to one or more abstract features. Moreover, while the representation of *intervention* remains unchanged, the remaining abstract features are represented as the count of characters holding that feature. This means that the distribution of harm between passengers and pedestrians and the legality of actions, which were characterized as "structural factors" by Awad et al. (2018), become attributes of the characters, who can be *passengers*, *pedestrians*, *law abiding* and *law violating*. In the example displayed in Figure 6, vector $\boldsymbol{\lambda}_1$ describes *intervention*, the presence of 2 *males*, 2 *females*, 3 *elderly* people, 4 *humans*, and 4 *law violating* individuals. A complete description of features before and after feature engineering is provided in Appendix A in Table 1 and Table 2 respectively.

This thesis replicates the extraction of abstract features performed by Kim et al. (2018) as laid down in their theoretical explanation. However, the feature engineering process employed is unique to this research. In particular, a function (*calculate_abstract_features*) is created which combines the information encoded in a given outcome $\boldsymbol{\theta}$ with that of the binary matrix representing the abstract feature space (Table 3). The function is used on the merged dataset representative of cultures around the world, in order to extract the 18 abstract features to be used as proxies for moral principles.

Figure 6: Example encoding



$$\theta_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad \lambda_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 0 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 0 \\ 4 \end{bmatrix}$$

Figure 6: From *Moral Machine*, n.d. by E. Awad, S. Dsouza, P. Chang. (www.moralmachine.net)[screenshot].

4.3 *Data science pipeline*

The research question proposed in this study investigates the influence of personal values on the decision to save of sacrifice characters in the MME. The target is the variable *saved*, whereas the abstract features extracted during the feature engineering step serve as predictor variables, and they are conceptualized as personal values for the purposes of this research. In order to investigate their influence on decisions, five sub-questions are formulated, which are reflected in the data science methodology described in the sections below. The methodology is divided into three parts: the analysis of increased dataset size (RQ 3), the analysis of model performance after hyperparameter tuning (RQ 1, 2 and 5), and the analysis of culturally homogeneous samples (RQ 4). The complete methodology and modeling schemata is outlined in Figure 7.

4.4 *Analysis of increased dataset size*

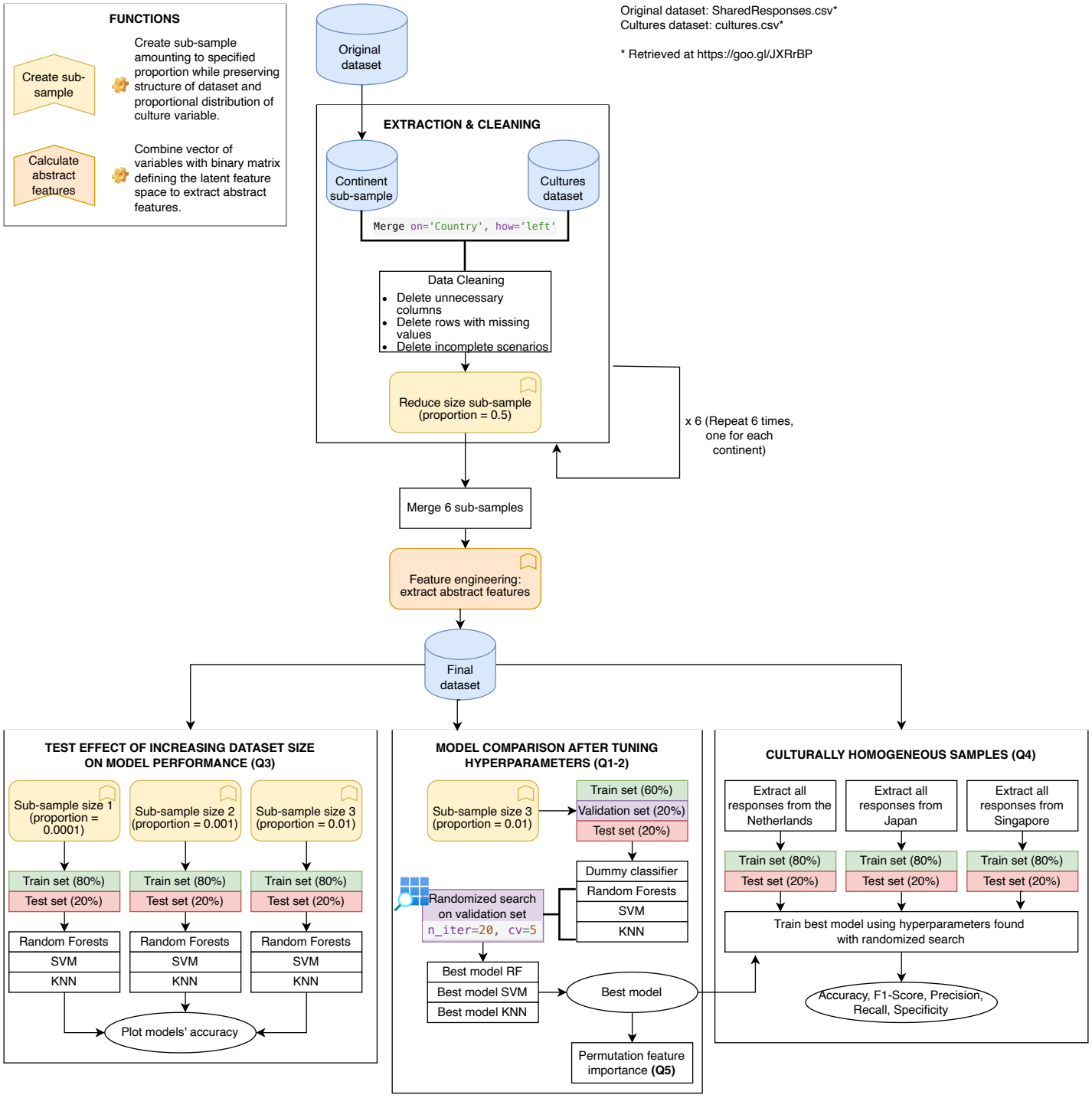
Sub-question 3 investigates the effect of increased dataset size on model performance. The literature suggests that understanding the relationship between samples size and model performance can shed light on the predictive capabilities of the selected independent variables (Agrawal et al., 2019; Cui & Gong, 2018; Kim et al., 2018). Kim et al. (2018) include as part of their benchmarking process a comparison of evaluation metrics across samples of increasing training set size. Agrawal et al. (2019) similarly compare various performance metrics including accuracy and area under the curve (AUC) across samples of increased dataset size. Finally, Cui and Gong (2018) do not study the MME, however, they find that when predicting individuals' behavior or cognitive abilities with various machine learning algorithms, exponentially increasing dataset size improves accuracy for all models and data types tested. Given these insights, three sub-samples are created by employing the *create_sub_sample* function described in the previous section. The proportions of the final dataset maintained for each sample are 0.0001, 0.001 and 0.01, amounting to 3,332, 33,310 and 333,100 rows respectively.

After separating the predictors from the target variable, each sample is subdivided into a training and testing set with 80-20 ratio to perform out-of-sample generalization. The size of the split was determined in part by replicating the methodology of Agrawal et al. (2019) and in part by considering the often invoked Pareto principle justification (Joseph, 2022). Finally, the Scikit-learn library is used to create an instance of a Random Forest Classifier, a Support Vector Classifier and a K-Neighbors Classifier, to be fit on each of the three samples of varying size. In this initial step, simple instances of the three classifiers are used, which do not provide hyperparameter specifications. Finally, classification accuracy is compared in order to make an initial assessment about the predictive potential of abstract feature.

4.5 *Analysis of model performance after hyperparameter tuning*

Sub-question 1 serves to specify the personal values employed as predictors, and coupled with sub-question 2 constitutes the most important part of this research. In particular, it inquires into the performance of the models when predicting decisions based on personal values. Given the computational requirements of tuning the models' hyperparameters by iteratively testing various combinations, a subset is extracted with the *create_sub_sample* function, amounting to 0.01 of the final dataset, or 333.100 rows. In this case, after separating the dependent and independent variables, the dataset

Figure 7: Flowchart of methodology and modeling



is split maintaining a 60-20-20 ratio, for training, hyperparameter tuning, and testing respectively. This ratio is commonly used in machine learning, especially when the dataset size is sufficiently large (Lever et al., 2016).

Similarly to the previous analysis, the three models are instantiated using the Scikit-learn library. In order to assess their performance before tuning the parameters, 5-fold cross validation is conducted on the training set, and the average validation score is considered. This validation method is chosen as it has been shown to perform well in striking a balance between bias and variance (Kohavi, 1995). The accuracy scores are used as reference to verify the effectiveness of the tuning process.

To assess the predictive capabilities of the models, their performance is compared to that of a naive classifier. Using a dummy classifier is helpful for setting a performance benchmark, since it highlights whether more complex models are able to learn additional useful patterns (Géron, 2022). As explained in the previous sections, the dataset used is designed to contain both sides of a scenario, which means that the classes of the target variable are perfectly balanced. Because of this, the dummy classifier with uniform strategy is selected from the Scikit-learn library. Uniform strategy is the most appropriate since it predicts the classes uniformly at random.

The process of hyperparameter tuning is performed on the validation set for the three models with random search cross-validation, or *RandomizedSearchCV* on Scikit-learn. This method was chosen instead of grid search cross-validation as it performs the search by randomly selecting parameters from a distribution, which substantially decreases the computational cost of the tuning operation, coupled with the computational cost of cross-validation, while maintaining its effectiveness (Bergstra & Bengio, 2012). The search is conducted over 20 iterations and 5 cross-validation folds, thus training a total of 100 models. This decision was made in light of the trade-off between computational efficiency and thoroughness of the search and considering the parameters most commonly used in existing literature (for example Anyanwu et al., 2023; Rimal et al., 2024).

The parameter grid for the Random Forest Classifier includes the number of decision trees in the random forest (search among values 100, 200 and 300), the maximum depth of each decision tree (search among values 10, 20 and no set limit), the minimum number of samples required to split an internal node (search among 2, 5 and 10) and to be at a leaf node (search among 1, 2, 4), as well as the amount of features to consider when searching for the best split (between the square root of the total number of features and the base-two logarithm of the total number of features). In particular, the amount of features considered controls the amount of randomness in the feature selection process, which means that smaller values such as base-two logarithm can reduce overfitting to the

training data (Breiman, 2001). Generally, other parameters which control the complexity of the decision trees within the random forest determine whether the model over- or underfits the training data (Hastie et al., 2009). For example, by not setting a maximum depth for the decision trees, the model might be able to capture more complex relationships in the data. Alternatively, when complexity leads to overfitting, the model might benefit from limited depth.

A similar search process was conducted for the Support Vector Classifier. The regularization parameter *C*, the kernel type, and the coefficient of the radial basis function (RBF) kernel coefficient *gamma* each contribute to the model's performance. The regularization parameter *C* helps the model to generalize well on unseen data while minimizing the training error (Cortes & Vapnik, 1995). Values between 10^{-2} and 10^{2} are used in the grid search, as the literature suggests that employing a broad initial range is effective for findings the correct balance (Hsu et al., 2003). Regarding the kernels parameter, linear and RBF kernel are included, as they are effective across various types of analysis (Schölkopf & Smola, 2002). Lastly, for the RBF kernel the *gamma* parameter determines the smoothness of the decision surface, so a broad range between 10^{-3} and 10^{2} is considered to ensure the thoroughness of the exploration (Hsu et al., 2003).

The hyperparameters considered for the K-Neighbors Classifier include the number of neighbors (*k*), distance metric, and weighting of neighbors. Given that the dataset contains over 300.000 rows and that the model employs 18 predictor variables, values searched for *k* range between 1 and 30. In fact, Cover and Hart (1967) note that the optimal value of *k* increases with the number of data points, so considering a broad range helps to smoothen the decision boundary without risking overfitting. The distance metrics searched include *Euclidean*, *Manhattan* and *Minkowski*, as optimizing the distance employed can improve the nearest neighbor error (Short & Fukunaga, 1981). Lastly, two weighing methods are searched, namely the *uniform* method which assigns equal influence to the *k* neighbors, and the *distance* method which assigns an influence inversely proportional to the distance from the query point. Testing these methods can enhance the model's ability to adapt to various data distributions (Mitchell, 1997).

After conducting random search cross-validation for the three models, the best performing models are selected and their performance metrics are compared, including accuracy, precision, recall, F1-score, specificity and AUC. Additionally, confusion matrices are created for each model, as they help visualize error patterns. Even though, these metrics can provide insights on the performance of the models, more importance is given to the accuracy values, given that the learning task involves a perfectly balanced target class, and that the outcome does not suffer

from obtaining predominantly more false positives or false negatives. Comparing accuracies, and considering the optimal hyperparameters, can helps answer the main research question by highlighting the predictive capabilities of personal values when using models with various algorithmic characteristics and in comparison to models used in existing literature.

The model with the highest accuracy is then selected to further study the influence of personal values. Particularly, to investigate sub-question 5, permutation feature importance (PFI) is performed with the best model, by using the *permutation_importance* callable from the *inspection* module of Scikit-learn. The method alters the values of each abstract feature analyzed, and inspects the relative drop in the model's performance (Breiman, 2001). This disruption in the feature-target relationship makes it possible to gauge the model's dependence on specific values when making predictions. In comparison to other feature importance measures such as SHAP values and LIME, PFI is relatively interpretable and computationally inexpensive. For example, SHAP values assess importance by contrasting predictions with and without specific features across all potential subsets, a process that becomes resource intensive due to the exponential proliferation of subsets (Lundberg & Lee, 2017). Similarly, LIME requires generating a large number of perturbed samples around each instance, as well as multiple runs (Ribeiro et al., 2016). Given these considerations, PFI is a suitable method for identifying the relative contribution of each personal value, which in turn sheds light on the most important predictors of decisions.

4.6 *Analysis of culturally homogeneous samples*

Sub-question 4 investigates whether model performance improves when predicting decision in culturally homogeneous samples. To answer the question, three samples are extracted from the final dataset corresponding to decisions made by respondents in the Netherlands, Japan and Singapore. The country samples were chosen as example cases, and they contain approximately the same amount of rows (254,554, 370,202 and 251,846 respectively). Importantly, their size is also similar to the size of the culturally diverse sample used to identify the best model. The three datasets are split into a training and test set with a ratio of 80-20, as hyperparameter tuning is not conducted for this portion of the analysis. Instead, the best model and corresponding best parameters previously identified with random search are used to produce performance metrics for the three country samples. The metrics, which include accuracy, precision, recall, F1-score and specificity, are compared with the metrics produced when fitting the culturally heterogeneous sample. Confusion matrices are also created for aiding comparison. This final portion of the analysis helps

to shed light on the predictive potential of personal values when influenced by cultural values.

4.7 *Language and packages*

The analysis was conducted using Python (version 3.9.13) programming language. The Pandas (version 1.4.4) and Numpy (version 1.26.4) libraries were especially useful for handling and manipulating large files in csv format (Harris et al., 2020; McKinney, 2010). Modeling was conducted with the Scikit-learn library (version 1.0.2) (Pedregosa et al., 2011). The use of callables *DummyClassifier*, *RandomForestClassifier*, *SVC*, *KNeighborsClassifier*, *RandomizedSearchCV*, and *permutation_importance* was especially useful. The visualization provided in the results section were produced with the aid of Matplotlib (version 3.5.2) and Seaborn (version 0.11.2) (Hunter, 2007; Waskom, 2021). Lastly, the world map in Figure 4 was created with the open source GeoPandas library (version 0.14.3) (Jordahl, 2016).

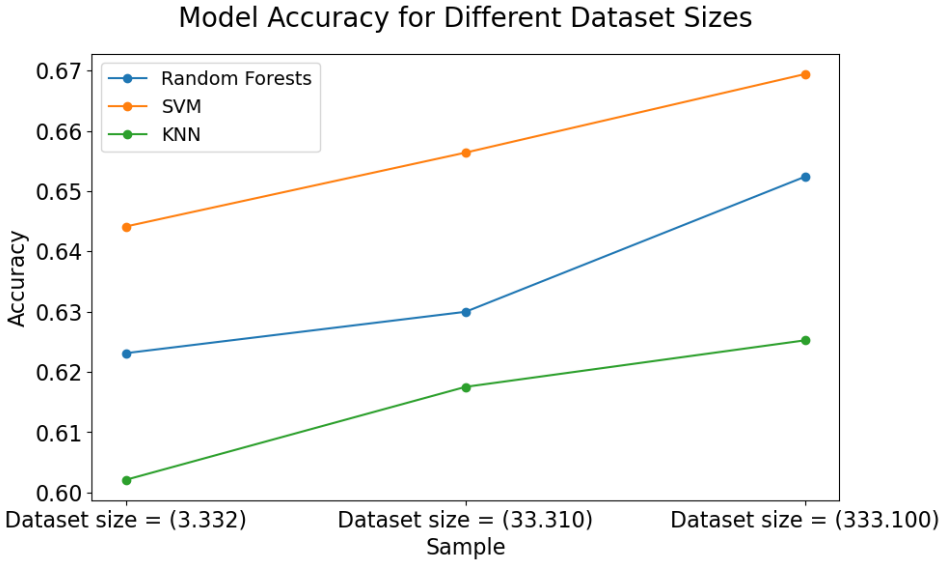
5 RESULTS

5.1 *Effect of increased sample size on model performance*

> **RQ3** *Does the predictive accuracy of personal values increase with increased sample size?*

The plot in Figure 8 displays the accuracy of the RF, SVM and KNN models before hyperparameter tuning across three sample sizes. The plot shows that as the dataset size increases, accuracy increases for all three models. This remains true within the tested range of roughly 3,000 to 300.000 observations. The plot also shows that SVM consistently outperforms random forests and KNN, with an accuracy ranging from 0.64 to 0.67. On the other hand, KNN produces the lowest accuracy value across the three samples, ranging between 0.60 and 0.62.

Figure 8: Performance comparison of models when increasing dataset size



5.2 *Model Comparison after hyperparameter-tuning*

RQ1 *Are personal values predictive of the decision to save or sacrifice characters in culturally diverse samples?*

RQ2 *Which of the following models is able to best predict decisions based on personal values: Random Forest, Support Vector Machine or K-Nearest Neighbors?*

Table 2 shows the average accuracy score after conducting 5-fold cross validation on the training set for the three models, before tuning hyperparameters. These values provide a baseline of comparison to verify whether the tuning process improves model performance. Similarly to the results in the previous section, SVM outperforms RF and KNN, while KNN produces the least accurate predictions.

Table 2: Model accuracy obtained with cross-validation on the training set before tuning hyperparameters

Metric	**RF**	**SVM**	**KNN**
Accuracy	0.65	0.67	0.63

Table 3 displays performance metrics of the dummy classifier, and the metrics of RF, SVM and KNN after tuning hyperparameters. The dummy

classifier is representative of models which do not learn patterns from the input data and randomly assign classes to the target variable. In fact, the performance metrics show that the naive model guesses about half of the correct classes. All three models obtain higher performance metrics when compared to the dummy classifier, suggesting that the chosen algorithms are able to use personal values to identify some patterns in relation to the target variable, and that they are not simply guessing.

Table 3: Performance metrics after tuning hyperparameters

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Specificity**
Dummy Classifier	0.50	0.50	0.50	0.50	0.50
Random Forests	0.66	0.66	0.68	0.67	0.65
SVM	0.67	0.67	0.66	0.67	0.68
KNN	0.66	0.66	0.65	0.66	0.67

The best RF model, which includes 200 decision trees, with a maximum depth of 10, improves over the original RF model by one point. The SVM model displays only a minor improvement in accuracy when using the optimal parameters found by the random search (from 0.6676 to 0.6684, rounded to 0.67 in the tables). The optimal model uses an RBF kernel and the regularization parameter C is set to 10. The KNN algorithm shows the greatest improvement in accuracy when using the optimal parameters, from 0.63 to 0.66. The best parameters found by the search use a k value of 29, the *Euclidean* distance and the *uniform* method. Even though SVM outperform RF in terms of accuracy, they obtain the same F1-score, which creates an harmonic mean for precision and recall. The F1-score of the KNN model remains slightly lower despite the improvement of the best model. Values for specificity highlight that SVM and KNN both outperform RF in correctly identifying the negative class. This is visible in the confusion matrices in Figure 9, 10 and 11. The RF model more often misclassifies the negative class where the respondent sacrifices the characters in the outcome, whereas SVM and KNN more often missclassify the positive class, where the respondent saves the characters in the outcome. As highlighted in the methodology section the best model is selected by reference to the accuracy score, since this research is interested in the overall amount of correct predictions, and considering that the target class is perfectly balanced. Therefore, SVM is selected as the best model.

Figure 9: RF - Confusion matrix with test set values

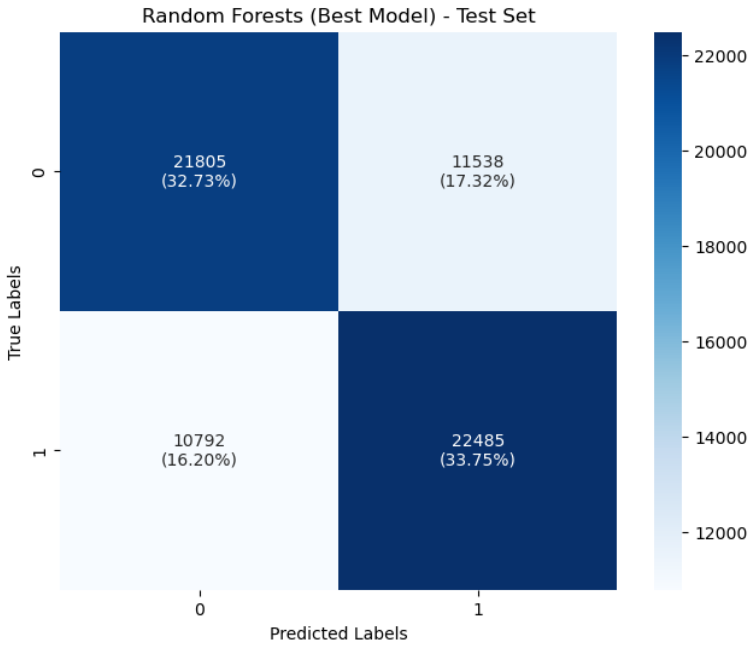


Figure 10: SVM - Confusion matrix with test set values

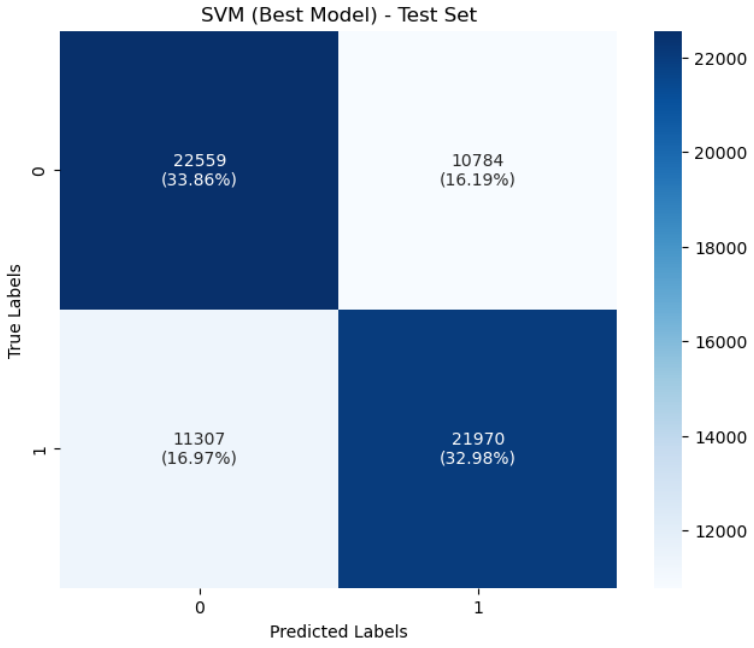


Figure 11: KNN - Confusion matrix with test set values

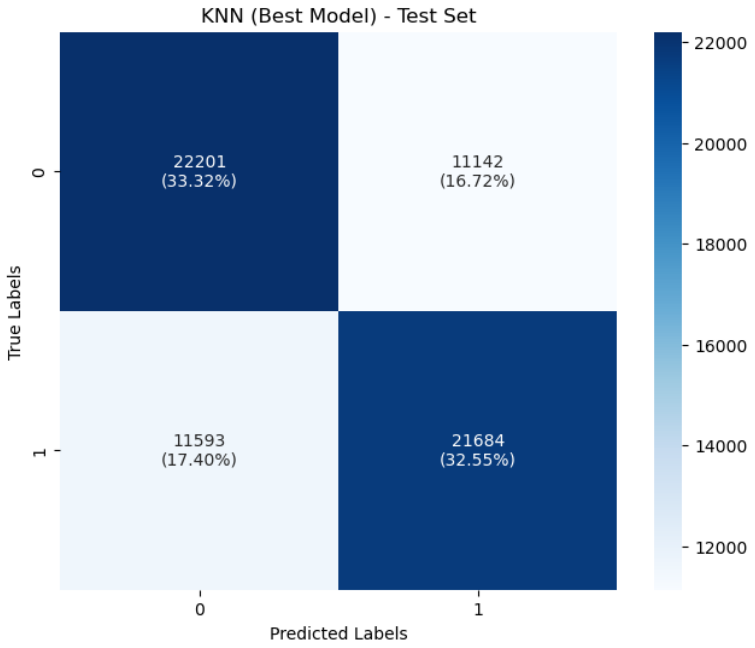
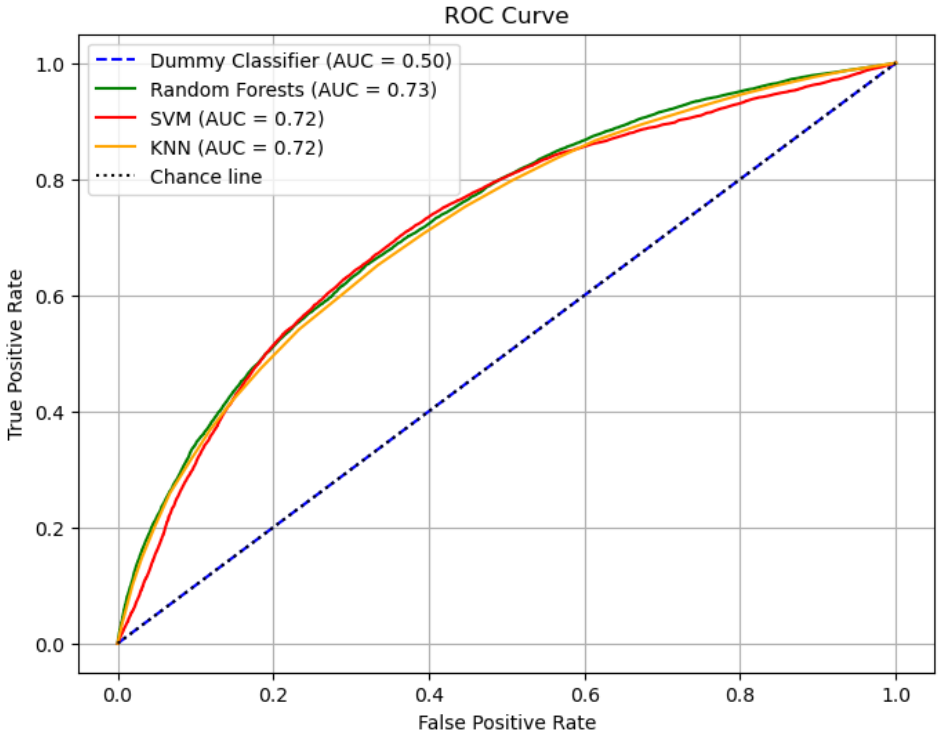


Figure 12 shows the Receiver Operating Characteristic (ROC) curve and AUC comparison of the models, which highlights the models' ability to distinguish between the positive and negative class. The dummy classifier lies on the chance line, while the three models display a similar AUC value. The AUC value for the RF model is slightly higher than the one for SVM and KNN. However, it is important to note that the AUC for SVM is calculated differently than for the other models. The AUC is calculated as the probability that each sample belongs to each class, however when instantiating the Support Vector Classifier in Scikit-learn, the parameter *probabilities* is set to *False*. Enabling it substantially increases the computational cost of the fitting process, since it makes use of 5-fold cross-validation to internally predict probabilities. Because of this, the *decision_function* method is used as a proxy, since it returns a score based on the distance of the sample from the decision boundary, which similarly represents whether the sample belongs to one class or the other.

Figure 12: ROC curve and AUC



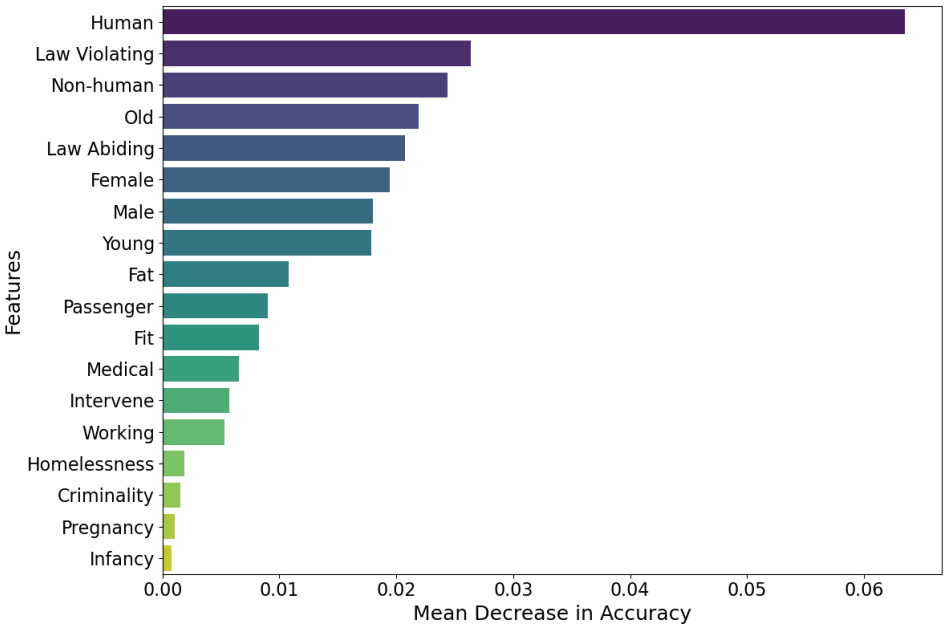
5.3 *Permutation feature importance of abstract features*

> **RQ5** *Which of the listed moral principles are more important for predicting decisions?*

The feature importance scores are calculated with the prevously selected best SVM model. These are presented in Figure 13, and are interpreted as the impact of each feature on the performance of the model. A positive high importance score indicates that shuffling its values leads to a decrease in the model's performance. The highest score is 0.063 for the feature *human*, while the feature *non-human* has a score of 0.024, which indicates that the model benefits from knowing whether a scenario contains humans as opposed to animals when making predictions. Other high ranking scores include *law violating* and *law abiding*, with a score of 0.026 and 0.020 respectively, and *old* with a score of 0.022. The features *female*, *male*, *young* and *fat* are also moderately high. The remaining features obtained a score lower than 0.01. These findings indicate that individuals place value on *species*, *the legality of actions*, *age*, *gender* and to a lesser degree *fitness* when making decisions. However, it should be noted that the feature

human is the most frequent feature in the scenarios, and that the four last features in Figure 13 are relatively uncommon, which indicates that feature importance scores are affected by how often a predictor variable is present. Nevertheless, the feature *non-human* is also relatively uncommon, thus indicating that *species* remains an important determinant for making decisions.

Figure 13: Permutation feature importance of abstract features



5.4 *Making predictions with a culturally homogeneous samples*

> **RQ4** *Is the predictive performance of values higher in culturally homogeneous samples?*

Table 4 shows the performance metrics obtained when training the best SVM model on three country samples sharing the same culture. All performance metrics displayed are slightly higher than the metrics for the culturally diverse sample.

Table 4: Performance metrics of SVM model on country samples

Country sampled	Accuracy	Precision	Recall	F1-Score	Specificity
Netherlands	0.68	0.68	0.68	0.68	0.68
Japan	0.67	0.68	0.67	0.67	0.70
Singapore	0.68	0.68	0.68	0.68	0.69

In particular, the model performs better on the samples from the Netherlands and Singapore, where the accuracy and F1-score reach a value of 0.68, an improvement over the performance on the culturally heterogeneous sample. In particular, the value for specificity is much higher in this case, indicating that the SVM algorithm is better able to predict the negative class. The confusion matrices in Figure 14, 15, and 16 show the count and proportion of observation within each true and predicted class. Comparing these figures with Figure 10, highlights that using the model on culturally homogeneous samples yields both a higher true positive (recall) and true negative rate (specificity).

Figure 14: The Netherlands - Confusion matrix

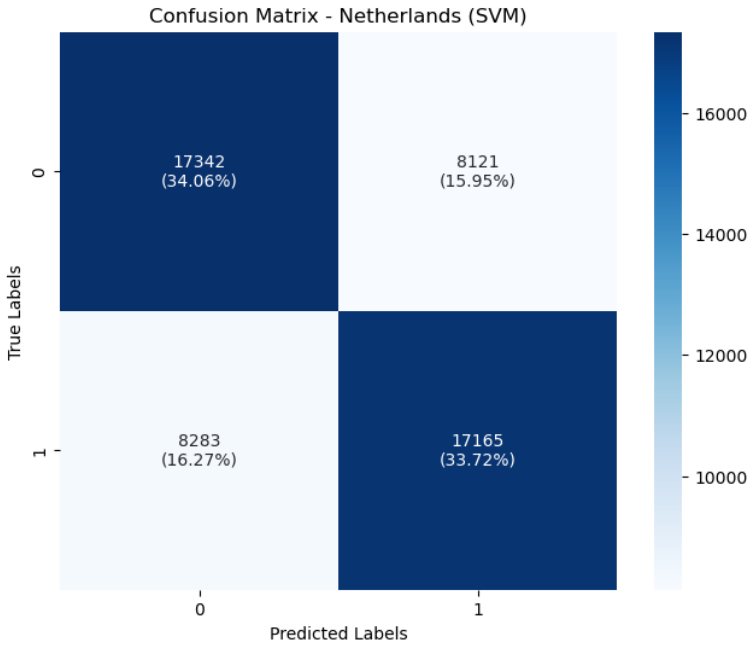


Figure 15: Japan - Confusion matrix

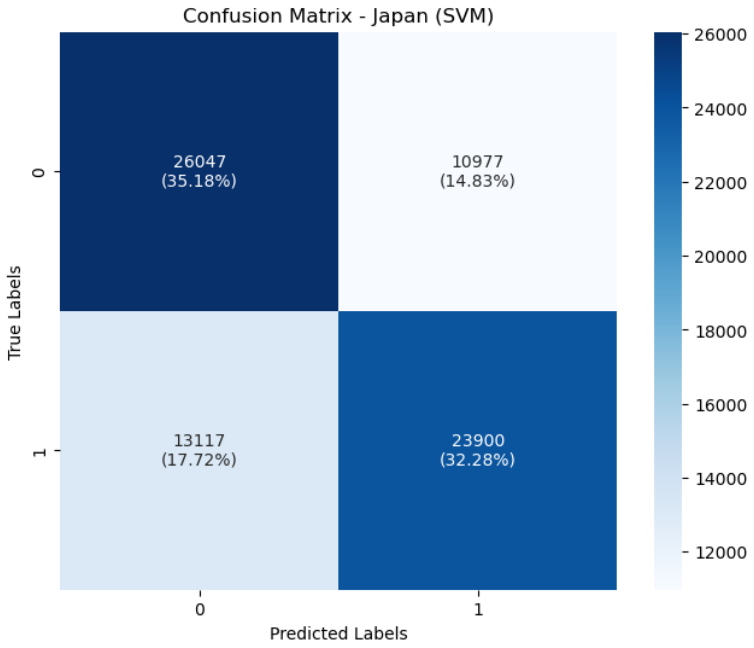
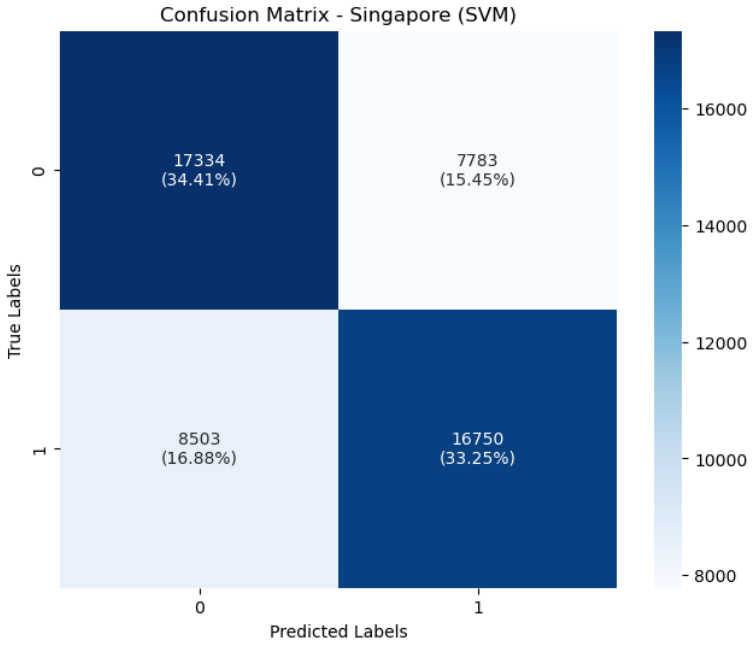


Figure 16: Singapore - Confusion matrix



6 DISCUSSION

The aim of this research is to uncover the influence of personal values on decision-making. Importantly, personal values are understood as independent from cultural values. To this end, a culturally heterogeneous sample is used to minimize reliance on cultural values when making predictions. Personal values are conceptualized as the relative importance given to each abstraction by the respondents. These abstractions are extracted from the variables by taking inspiration from the theoretical model described by Kim et al. (2018), where each abstract component belongs to one or more characters. Existing literature suggests that abstractions are a fundamental instrument that helps human moral reasoning to create value hierarchies (Blasi, 1980; Kohlberg, 1981), and that trolley-like dilemmas are able to accentuate reasoning processes (Luft, 2020; Plunkett & Greene, 2019). This thesis explores the predictive potential of personal values when using RF, SVM and KNN, as these fundamental machine learning algorithms have been used in existing studies modelling human behavior and decision-making, but they have not been employed in the context of the MME (Bourgin et al., 2019; Plonsky et al., 2017, 2019).

The findings suggest that the predictive accuracy of personal values increases with increased sample size (RQ3), which is in line with expectations. The models employed by Agrawal et al. (2019) display a similar behavior, more specifically when increasing the size of the dataset up to 10^5 observations, as similarly done here. This highlights that large collections of human value preferences can be exploited to devise models which are able to predict human decision-making. Collecting more data like the one provided by the MME, including relating to other types of decision-making, will serve to improve machines' capability to mimic human choices and align these choices more closely to human values. For example, Sierra et al. (2021) integrate values in their formal approach for developing autonomous systems, in order to ensure value alignment. Thus, the study of values benefits from extensive data collection and analysis, such as the one conducted in this research.

The models are trained on a culturally heterogeneous sample containing responses from all continents and including all nine cultures of the Inglehart-Welzel Cultural Map (Inglehart & Welzel, 2005). This approach is used to minimize the influence of shared cultural values to make predictions, while maximizing the focus on personal values. The results show that SVM yields the highest value for accuracy (0.67), however, RF yields a similar F1-score (0.67), and the performance of KNN improves considerably after hyperparameter tuning (RQ2). All three models clearly outperform the naive classifier, thus highlighting that they are able to capture some

patterns in the relationship between personal values and respondents' moral decisions (RQ1). However, the benchmark accuracies presented in the existing literature surpass the accuracies obtained here. Particularly, neural networks are able to reach an accuracy value of 0.80 (Agrawal et al., 2019), and hierarchical structures which assign weights to features are able to reach an accuracy value of 0.75 (Kim et al., 2018). Because of this, future research should focus on developing models which better calibrate the contribution of individual values, for example by combining classic machine learning algorithms with models that assign weights to predictor variables, as done by Plonsky et al. (2019). When comparing the performance metrics for culturally heterogeneous samples with those of homogeneous samples, the results of the latter are consistently higher (RQ4). This finding helps to solidify the assumptions made about shared cultural values, and it emphasizes the importance of studying values in relation to culture.

An additional contribution of this research relates to the relative importance of personal value in determining decisions. To this aim, permutation feature importance is performed on the best performing SVM model. The results suggest that the model predominantly relies on *species*, *legality* and *age* to predict whether the characters in an outcome are saved or sacrificed (RQ5). Despite this, it should be noted that the obtained feature importance scores are partially related to the frequency with which the various features appear in the scenarios. Nevertheless, it can be argued with sufficient certainty that important considerations for respondents include whether humans or animals are involved, whether the victims are young or old and whether an action is legal. In particular, the former two considerations are in line with the findings of the original experiment by Awad et al. (2018). Considering the importance of factors such as *species*, *legality* and *age* can serve as a starting point for discussions on the value alignment objectives of AVs. However, these results are more immediately related to the predictive process used by the SVM model, instead of the reasoning process employed by individuals. Individuals' moral reasoning and values are more easily brought forth by qualitative investigations such as the one conducted by Rhim (2020). As pointed out by LaCroix (2022), machine learning benchmarks such as the ones created by Kim et al. (2018) cannot create a threshold for machine ethical decision-making, because models simply represent respondents' understanding of morality, they do not define morality per se. When working on value alignment in the context of AVs, an adaptive approach to values should be preferred to a one-size-fits-all approach, as similarly argued by Nallur (2020) and (Arnold et al., 2017).

6.1 *Contribution, limitations and future direction*

This research contributes to the literature bridging values and machine learning, by studying the influence of personal values on moral judgement by means of three underused machine learning algorithms in the context of the MME. Additionally, it focuses on a heterogenous sample and extracts the feature importance measure of each variable. Overall, these results contribute to the academic discussion on value alignment in the context of AV moral decision-making.

Nevertheless, the study has some limitations, as the elements which are being measured are simply proxies of real phenomena and characteristics. For example, it relies on a culturally diverse sample of respondents, however, assigning one single culture to a country is diminutive of within-country variations. Respondents are only assumed to belong to a certain culture. Additionally, variations in moral judgement could be determined by factors which are not culturally or territorially determined, such as for example, demographic factors, income, education, political affiliation and exposure to internet content. The extracted abstract features are also proxies for personal values, since the actual moral reasoning of individuals can only be discovered through qualitative discussion, as conducted by Rhim (2020). Finally, responses are taken to represent proxies for moral facts, as further highlighted by LaCroix (2022).

Future research should focus on combining existing machine learning algorithms with models that can assign different weights to features. Collecting greater quantities of behavioral data would also aid the capabilities of models developed to predict human behavior and decision-making, while ingraining human values within prediction processes. These models can in turn be used to create more adaptive algorithms steering the functioning of AVs. Finally, more attention should be paid to developing a better qualitative understanding of personal values and moral reasoning across various groups, which can lay the foundation for more accurate quantitative models.

7 CONCLUSION

This research investigates the influence of personal values on decision-making in the moral dilemma scenarios of the MME. Three classic machine learning algorithms, RF, SVM and KNN are used to make predictions on culturally diverse responses, in order to better isolate personal from cultural values, given the importance of their separation according to value scholars (Hofstede, 1984; Schwartz, 2006). When examining three samples of increasing size, the predictive accuracy of the models increases,

suggesting that noisy and diverse human behavior can be better modeled with extensive data. When examining a culturally diverse sample of response, SVM shows the highest accuracy after hyperparameter tuning, even though, all models outperform the naive classifier. This highlights that all three models are able to identify patterns in the relationship between personal values and moral decisions. Future studies should further develop these models by introducing theoretical priors, in order to improve the prediction accuracy of moral decisions. Overall, this study underscores the interplay between personal and cultural values, suggesting that future research should integrate both qualitative and quantitative approaches to better distinguish the two and to better model human moral reasoning based on a priori assumptions. In particular, the AV industry would benefit from such models, as existing research focuses on introducing values as part of systems' autonomous decision-making, with the aim of solving the value alignment problem (Peterson, 2019; Sierra et al., 2021).

REFERENCES

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2019). Using machine learning to guide cognitive modeling: A case study in moral reasoning. *arXiv preprint arXiv:1902.06744*.

Agresti, A. (2012). *Categorical data analysis (vol. 792)*. John Wiley & Sons.

Anyanwu, G. O., Nwakanma, C. I., Lee, J.-M., & Kim, D.-S. (2023). Falsification detection system for iov using randomized search optimization ensemble algorithm. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 4158–4172.

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment–what will keep systems accountable? *Workshops at the thirty-first AAAI conference on artificial intelligence*.

Awad, E., Dsouza, S., & Chang, P. (n.d.). *Moral machine*. www.moralmachine.net

Awad, E., Dsouza, S., Kim, R., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85.

Berberich, N., & Diepold, K. (2018). The virtuous machine - old ethics for new technology? *arXiv*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological bulletin*, *88*(1), 1–45.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, *29*(7), 1084–1093.

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. *International conference on machine learning*, 5133–5141.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A., & Yampolskiy, R. (2017). Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*.

Christian, B. (2021). *The alignment problem: How can machines learn human values?* Atlantic Books.

Conitzer, V., Sinnott-Armstrong, W., Schaich Borg, J., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), 4831–4835.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage*, *178*, 622–63.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*, 139–157.

Elster, A., & Gelfand, M. J. (2021). When guiding principles do not guide: The moderating effects of cultural tightness on value-behavior links. *Journal of Personality*, *89*(2), 325–337.

Feather, N. (1988). Moral judgement and human values. *British Journal of Social Psychology*, *27*(3), 239–246.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25–42.

Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow*. O'Reilly Media, Inc.

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2424*(1), 58–65.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Hadfield-Menell, D., & Hadfield, G. K. (2019). Incomplete contracting and ai alignment. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 417–422.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5). sage.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence (vol. 333)*. Cambridge: Cambridge university press.

Inglehart, R. (2020). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton university press.

Jordahl, K. (2016). Geopandas documentation. *URL: http://sethc23. github. io/wiki/Python/geopandas. pdf-Download vom*, *26*, 2022.

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J., & Rahwan, I. (2018). A computational model of commonsense moral decision making. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 197–2013.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

Kohlberg, L. (1981). *The philosophy of moral development: Moral stages and the idea of justice (vol. 1)*. Harper & Row.

LaCroix, T. (2022). Moral dilemmas for moral machines. *AI and Ethics*, *2*(4), 737–746.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: Model selection and overfitting. *Nature methods*, *13*(9), 703–705.

Luft, A. (2020). Theorizing moral cognition: Culture in action, situations, and relationships. *Socius*, *6*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

McKinney, W. (2010). Data structures for statistical computing in python. *SciPy*, *445*(1), 51–56.

Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect. *Georgetown Public Law Research Paper*, *762385*.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

Mitchell, T. M. (1997). Instance-based learning [A comprehensive introduction to the field of machine learning, providing insights into various algorithms and methodologies.]. In *Machine learning* (pp. 230–267). McGraw Hill.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18–21.

Nallur, V. (2020). Landscape of machine implemented ethics. *Science and engineering ethics*, *26*(5), 2381–2399.

Noothigattu, R., Gaikwad, S. '. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. D. (2018). A voting-based system for ethical decision making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Ostini, R., & Ellerman, D. A. (1997). Clarifying the relationship between values and moral judgement. *Psychological reports*, *81*(2), 691–702.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). The role of analytic thinking in moral judgements and values. *Thinking & reasoning*, *20*(2), 188–214.

Peterson, M. (2019). The value alignment problem: A geometric approach. *Ethics and Information Technology*, *21*, 19–28.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., & Carter, E. C. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on bostyn, sevenhant, and roets (2018). *Psychological Science*, *30*(9), 1389–1391.

Ravlin, E. C., & Meglino, B. M. (1987). Effect of values on perception and decision making: A study of alternative work values measures. *Journal of Applied psychology*, *72*(4), 666.

Rest, J. R. (1986). *Moral development: Advances in research and theory*. New York: Praeger.

Rhim, J. a. (2020). Human moral reasoning types in autonomous vehicle moral dilemma: A cross-cultural comparison of korea and canada. *Computers in Human Behavior*, *102*, 39–56.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rimal, Y., Sharma, N., & Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: Student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimedia Tools and Applications*, 1–16.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351–357.

Rokeach, M. (1973). *The nature of human values.* Free press.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Sagiv, L., & Schwartz, S. H. (2022). Personal values across cultures. *Annual review of psychology*, *73*, 517–546.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Schwartz, S. (2006). A theory of cultural value orientations: Explication and applications. *Comparative sociology*, *5*(2-3), 137–182.

Seifert, J., Friedrich, O., & Schleidgen, S. (2022). Imitating the human. new human–machine interactions in social robots. *NanoEthics*, *16*(2), 181–192.

Short, R., & Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification. *IEEE transactions on Information Theory*, *27*(5), 622–627.

Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., & Perelló, A. (2021). Value alignment: A formal approach. *arXiv preprint arXiv:2110.09240*.

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, *10*(2), 124–140.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–2017.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Weber, J. (1993). Exploring the relationship between personal values and moral reasoning. *Human Relations*, *46*(4), 435–463.

Wiedeman, C., Wang, G., & Kruger, U. (2020). Modeling of moral decisions with deep learning. *Visual Computing for Industry, Biomedicine, and Art*, *3*(27).

Wiener, N. (1960). Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, *131*(3410), 1355–1358.

WVSA. (2020). *World values survey*. www.worldvaluessurvey.org

APPENDIX A

Table 1: Description of variables before feature engineering

Variable name	Description	Type
ResponseID	Unique, random set of characters that represents an identifier of the scenario. Two rows share the same ResponseID	Categorical
UserCountry3	Alpha-3 ISO code of the country from which the user accessed the website	Categorical
Culture	Culture assigned to the respective country by the Inglehart-Welzel Cultural Map 2010-2014	Categorical
Intervention	Decision of the AV to stay (0) or swerve (1)	Integer
Barrier	Indicates that the potential casualties in this outcome are passengers (1) or pedestrians (0)	Integer
CrossingSignal	Indicates that there is no traffic light in the outcome (0), that there is a green traffic light (1), or that there is a red traffic light (2)	Integer
NumberOfCharacters	Takes a value between 1 and 5, indicating the total number of characters in the outcome	Integer
Man	General man (count between 0 and 5)	Integer
Woman	General woman (count between 0 and 5)	Integer
Pregnant	A visibly pregnant woman (count between 0 and 5)	Integer
Stroller	Stroller with a baby in it (count between 0 and 5)	Integer
OldMan	White-haired old man walking with a stick (count between 0 and 5)	Integer
OldWoman	White-haired old woman walking with a stick (count between 0 and 5)	Integer
Boy	Young boy, child (count between 0 and 5)	Integer
Girl	Young girl, child (count between 0 and 5)	Integer
Homeless	Man with beard and ragged attire (count between 0 and 5)	Integer
LargeWoman	Large woman (count between 0 and 5)	Integer
LargeMan	Large man (count between 0 and 5)	Integer

Continued on next page

Table 1: Continued from previous page

Variable name	Description	Type
Criminal	Man with a mask stealing a bag of money (count between 0 and 5)	Integer
MaleExecutive	Working man with suit and briefcase (count between 0 and 5)	Integer
MaleAthlete	Man jogging in shorts (count between 0 and 5)	Integer
MaleDoctor	Man holding an emergency kit (count between 0 and 5)	Integer
FemaleExecutive	Working woman with suit and briefcase (count between 0 and 5)	Integer
FemaleAthlete	Woman jogging in shorts (count between 0 and 5)	Integer
FemaleDoctor	Woman holding an emergency kit (count between 0 and 5)	Integer
Dog	Dog (count between 0 and 5)	Integer
Cat	Cat (count between 0 and 5)	Integer
Saved (target)	User saved the characters in the outcome (1), or user sacrificed the characters in the outcome (0)	Integer

Table 2: Description of variables after feature engineering

Variable name	Description	Type
ResponseID	Unique, random set of characters that represents an identifier of the scenario. Two rows share the same ResponseID	Categorical
UserCountry3	Alpha-3 ISO code of the country from which the user accessed the website	Categorical
Culture	Culture assigned to the respective country by the Inglehart-Welzel Cultural Map 2010-2014	Categorical
Intervene	Decision of the AV to stay (0) or swerve(1)	Integer
Male	Number of male characters (count between 0 and 5), including *Man*, *OldMan*, *Boy*, *LargeMan*, *MaleExecutive*, *MaleAthlete* and *MaleDoctor*	Integer

Continued on next page

Table 2: Continued from previous page

Variable name	Description	Type
Female	Number of female characters (count between 0 and 5), including *Woman*, *Pregnant*, *OldWoman*, *Girl*, *LargeWoman*, *FemaleExecutive*, *FemaleAthlete* and *FemaleDoctor*	*Integer*
Young	Number of children (count between 0 and 5), including *Boy*, *Girl* and *Stroller*	Integer
Old	Number of old characters (count between 0 and 5), including *OldMan* and *OldWoman*	Integer
Infancy	Number of infant characters (count between 0 and 5), including *Stroller*	Integer
Pregnancy	Number of pregnant characters (count between 0 and 5), including *Pregnant*	Integer
Fat	Number of large characters (count between 0 and 5), including *LargeMan* and *LargeWoman*	Integer
Fit	Number of fit characters (count between 0 and 5), including *MaleAthlete* and *FemaleAthlete*	Integer
Working	Number of executives (count between 0 and 5), including *MaleExecutive* and *FemaleExecutive*	Integer
Medical	Number of medical professionals (count between 0 and 5), including *MaleDoctor* and *FemaleDoctor*	Integer
Homelessness	Number of characters depicted as homeless (count between 0 and 5), including *Homeless*	Integer
Criminality	Number of criminals (count between 0 and 5), including *Criminal*	Integer
Human	Number of human characters (count between 0 and 5), including *Man*, *Woman*, *Pregnant*, *Stroller*, *OldMan*, *OldWoman*, *Boy*, *Girl*, *Homeless*, *LargeWoman*, *LargeMan*, *Criminal*, *MaleExecutive*, *FemaleExecutive*, *FemaleAthlete*, *MaleAthlete*, *FemaleDoctor* and *MaleDoctor*	Integer
Non-human	Number of animals (count between 0 and 5), including *Dog*, and *Cat*	Integer

Continued on next page

Table 2: Continued from previous page

Variable name	Description	Type
Passenger	Number of passengers in the AV (count between 0 and 5)	Integer
Law abiding	Number of pedestrians crossing with a green light (count between 0 and 5)	Integer
Law violating	Number of pedestrians crossing with a red light (count between 0 and 5)	Integer
Saved (target)	User saved the characters in the outcome (1), or user sacrificed the characters in the outcome (0)	Integer