DETECTING THE USE OF LARGE LANGUAGE MODELS IN THE MORAL MACHINE EXPERIMENT

COMPARING THE PERFORMANCES OF LOGISTIC REGRESSION, RANDOM FOREST, SUPPORT VECTOR MACHINE AND MULTILAYER PERCEPTRON

EMMA KUIPERS

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

u857423

COMMITTEE

dr. Michał Klincewicz
Mr. M. Zamanzadeh Nasrabadi

LOCATION

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

DATE

December 2nd, 2024

WORD COUNT

8796

ACKNOWLEDGMENTS

I would like to thank my supervisor dr Michał Klincewicz for his help throughout the thesis writing process. Without his tips, words of encouragement and questions that made me critically reflect on my work, I would not have been able to submit the work that is in front of you right now. On top of that, my gratitude goes out to my fellow thesis writing students, who I was very grateful for to consult with and who understood what I was going through. I would like to thank my family who provided me with a safe haven where I had all the space, time and peace to complete this project. Finally, I could not have survived the long days in the library without some of my closest friends there with me. They have gotten me through the rough patches of this thesis project by providing me with new useful insights, being there to listen to and help me think through some of the problems I encountered and by just making me laugh when I had a tough time. I will forever be grateful for your help and support in these days.

DETECTING THE USE OF LARGE LANGUAGE MODELS IN THE MORAL MACHINE EXPERIMENT

COMPARING THE PERFORMANCES OF LOGISTIC REGRESSION, RANDOM FOREST, SUPPORT VECTOR MACHINE AND MULTILAYER PERCEPTRON

EMMA KUIPERS

Abstract

This study examines the ability of Machine Learning (ML) models to effectively differentiate whether responses to Moral Machine Experiment (MME, Awad et al. (2018)) scenarios were generated by Large Language Models (LLMs) or humans. Four models—Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—were evaluated across three datasets: one including respondents from all ethnicities (df_total), one limited to non-Western respondents (df_nw), and one with underrepresented values removed (df_deleted). Error analysis, including feature importance, was conducted to explore if and how moral frameworks of humans and LLMs might differ.

Results show that the RF model consistently outperformed the others, achieving macro F1-scores between 0.685 and 0.713 and Recall scores for class 1 around 0.672 on the test set, highlighting its robustness and reliability. No significant differences were observed between the error patterns of the df_total and df_nw datasets, suggesting minimal divergence in moral frameworks across cultural backgrounds. Despite slightly lower F1-scores, the RF model on the df_deleted dataset demonstrated adaptability, achieving similar Recall scores while avoiding reliance on the shortcuts present in the other datasets. Error analysis revealed utilitarian dilemmas as a key area of divergence between human and LLM moral frameworks, with features representing utilitarian dilemmas being the most important.

These findings contribute to scientific understanding of ML models' performance in distinguishing between LLM and human moral decisions and underscore the areas where LLMs still differ from humans. The results also provide actionable insights for policymakers addressing LLM detection in moral decision-making contexts.

0 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

This research uses data from the Moral Machine Experiment, which is a dataset created by (Awad et al., 2018). This data was obtained from OSF and can be downloaded here. Moreover, several datasets created and managed by (Takemoto, 2024) were utilized in this study. These datasets are publicly available via this link. The original creators of these datasets remain ownership of their data after the completion of this research. By making the data publicly available, both of these authors have given consent for the use and manipulation of the data for research purposes. All information present in the data used in this study was anonymized. No where in this thesis was it necessary to collect data from humans or animals. All figures, except for Figure 1, have been created by the author of this thesis. Code that was used to generate the results published in this study can be found in the 'Moral-Machine-2024' repository of dr. Klincewicz, under Emmakuipers.The different libraries, programs and packages used in this research can be found in Appendix A (page 45). A generative language model (OpenAI's ChatGPT-4.0) has been used to improve the authors original content, and for paraphrasing, code debugging and grammar. Overleaf in combination with LaTex were used to format this research. No other technological tools or services were utilized.

1 INTRODUCTION

As Artificial Intelligence (AI) is becoming more prominent within current society, the accompanying legislation becomes as well (Norden & Lerude, 2023). Different unions and regions worldwide have established extensive regulations regarding AI use (European Parliament, 2024; Association of Southeast Asian Nations, 2024; African Union, 2024). These policies state the importance of transparency when using AI, especially when AI use is considered 'high risk'. High risk AI systems have a high probability of negatively impacting the safety of their environment and/or are systems that have high severity of harm (European Parliament, 2023; Association of Southeast Asian Nations, 2024). Due to the serious consequences that may arise if issues occur in these areas, it is of utmost importance that these strict rules of transparency are adhered to.

An area in which high-risk AI systems are used is in the automotive industry (European Parliament, 2023), particularly when designing autonomous vehicles (AV). Here, AI systems like Large Language Models (LLMs), might act as moral decision-makers in situations of life and death. This could become reality rather sooner than later, as the automotive in-

dustry anticipates on incorporating LLMs to assist in AVs decision-making processes (Lei et al., 2023). They may, for example, be employed to decide whether to prioritize the safety of its passengers or pedestrians in case of vehicle malfunction. The disastrous consequences of non-disclosure, along with the possibility that (accidentally or on purpose) the use of LLMs is not reported by the programmer, highlight the need for stricter oversight of the use of LLMs in the AV industry. This could be achieved by developing a model that can detect whether or not LLMs were used in moral decision-making. Machine Learning (ML) models in particular suit this task as they have proven to be successful in previous detection tasks (Crawford et al., 2015). The main goal of this research is therefore assessing whether and how effectively ML models can detect the use of LLMs in data containing moral decision-making in the AV industry.

The Moral Machine Experiment (MME) from Awad et al. (2018) provides this data on moral decision-making in the AV industry. The MME is an online experiment that gathered data on the decisions humans made when faced with moral dilemma's concerning AV's. Takemoto (2024) build on this by repeating the experiment, but taking LLMs as respondents instead of humans. Current research combines data from both studies to achieve the goal of this research: trying to, with help of ML models, accurately detect the use of LLMs in moral decision-making in the AV industry.

The difference in data collection between the studies created a divergence between the scenarios presented to the LLMs compared to the humans respondents. This unfair distribution of scenarios might result in the ML models learning unintended associations, instead of identifying the LLMs based on their moral decision-making. This research will therefore also investigate the potential influence of these dataset differences. Section 3.2.2 outlines the methodology for this analysis.

Moreover, Takemoto (2024) found that LLMs and humans are generally aligned in their moral decisions, but stated that cultural and societal factors were not considered. They suspect however that these factors, which are critical in forming moral preferences (Liu et al., 2024), might bias LLMs to lean towards moral decisions in line with Western culture (Takemoto, 2024). This research addresses this by examining the possible influence of cultural factors on the performance of the models.

While research suggests that while LLMs and humans largely share similar moral frameworks, some have found that certain LLMs deviate from human preferences in specific MME scenarios (Takemoto, 2024; (Vida et al., 2024)). To asses the extent to which this holds true in current re-

search, and for which moral scenario types this might be the case, feature importance and error analysis techniques are employed. These identify key features and scenario types, highlighting where moral decisions differ most between humans and LLMs.

Summarizing, no prior research has explored the detection of LLM usage in moral decision-making within the AV industry. This study pioneers by being the first to apply ML models for this task, thereby filling this gap in the literature.

1.1 *Research Questions*

The main research question that flows from the above is:

> *To what extent can various machine learning models effectively detect whether a response to a Moral Machine Experiment scenario was generated by a Large Language Model or by a human, in different contexts?*

The sub-questions that will help answering this main question are:

SRQ1 *In what way do the performance metrics of a Random Forest, Support Vector Machine and a Multilayer Perceptron model differ when asked to detect whether a response to a Moral Machine Experiment scenario was generated by a Large Language Model or by a human, and which of these models has the highest performance metrics in this task?*

SRQ2 *Does the predictive performance of the models change when accounting for the discrepancies between the scenarios presented to LLMs and humans? And if so, how?*

SRQ3 *How are the model performances affected when considering the ethnicity of human respondents, and what does this say about the alignment of the moral frameworks of humans and LLMs?*

SRQ4 *Which moral scenario types are most important in distinguishing between human and LLM-generated moral decisions, and how do these types influence the classifications made by the best-performing model(s)*

1.2 *Summary of findings*

This study identifies the Random Forest (RF) model as the most effective for detecting LLM-generated responses in the MME dataset, consistently outperforming other models with Recall scores for class 1 ranging from 0.672

to 0.673. Analysis moreover shows that removing scenario inconsistencies had minimal impact on performance and suggests limited differences between the moral frameworks of non-Western respondents and LLMs and Western and non-Western respondents. Utilitarian scenarios were found to be the area where humans and LLMs differ the most in their moral standpoint, with features from this scenario type being the most important across datasets.

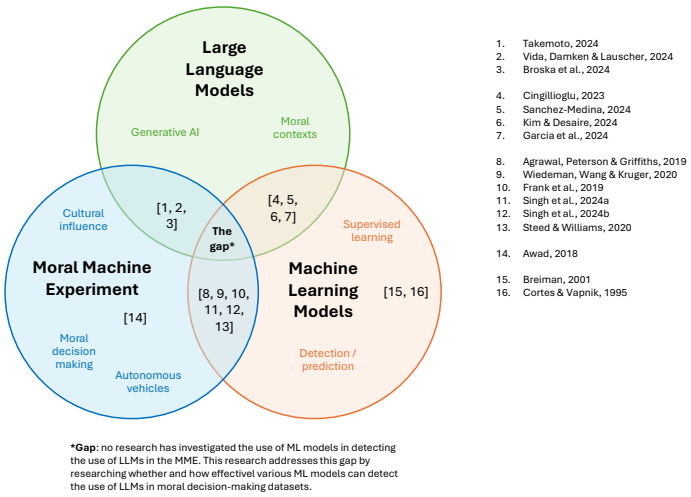


Figure 1: Literature gap indicated by means of a Venn diagram

2 RELATED WORK

The literature gap that this research aims to fill is presented in Figure 1.

2.1 *Moral decision-making of humans and LLMs (RQ & SRQ4)*

Research on moral decision-making dates back to the trolley problem dilemma of Foot (1967), which started a scientific discourse on how human ethical frameworks shape moral choices. Building on this foundation, Awad et al. (2018) introduced the Moral Machine Experiment (MME). This is an online game-like experiment in the context of Autonomic Vehicles (AV's), presenting moral dilemmas where respondents had to choose between two outcomes. Each scenario is characterized by a moral scenario types (see Table 1, and their research uncovered how cultural and demographic differences impact ethical preferences (Awad et al., 2018).

Table 1 shows the found global moral preferences of over 35 million users that participated in the MME. The table shows that humans prefer to save the characters with attributes present in the bold text, rather than the other option. For example, respondents have a preference for saving more characters over less (Utilitarian Scenario Type), or females over males (Species Scenario Type).

The emergence of LLMs has raised interest in how these models make moral decisions, especially in comparison to humans. Takemoto (2024) was the first to investigate ethical decision-making tendencies of LLMs

Scenario Type	Choice
Gender	Sparing males – **Sparing females**
Fitness	Sparing the large – **Sparing the Fit**
Social Status	Sparing lower social status – **Sparing higher social status**
Age	**Sparing younger** – Sparing Older
No. Characters	Sparing fewer characters – **Sparing more characters**
Species	**Sparing humans** – Sparing pets

Table 1: Global preferences of humans per moral Scenario Type, denoted in bold, in the MME. Source: Awad et al. (2018)

in the context of the MME, finding substantial differences between human and LLM-generated moral decisions. The LLM 'PaLM 2' showed the most pronounced difference from human preferences found by Awad et al. (2018), in that it preferred to save less people over more, and less fit people over fitter humans (Takemoto, 2024). Vida et al. (2024) expanded on this by exploring moral preferences of LLMs in a multilingual setting and comparing those to humans. Their study revealed that moral biases do exist, to some extent, within LLMs, and that their moral preferences differ, sometimes greatly, from human preferences in certain scenario types (Vida et al., 2024). An example is that the 'Llama 3 70B-Instruct' LLM tends to run over more rather than fewer people when given the choice, as well as rather running over humans than pets.

The main purpose of this study (*RQ*) is assessing whether ML models can effectively detect LLM from human responses to moral dilemmas. The findings above suggest that significant disparities still exist between ethical frameworks of humans and LLMs in the AV context. Consequently, this research hypothesizes that ML models can identify these differences, thereby classifying whether responses to MME scenarios were human or LLM generated.

Additionally, as was found that LLMs diverge from humans in certain types of moral scenarios, especially in scenarios with utilitarian, species and fitness attributes, this research expects that these scenario types are of great importance for classifications made by the best-performing model(s) (*SRQ4*).

2.1.1 *Culturally influenced ethical frameworks (SRQ3)*

Ethical frameworks, on which moral decisions are based, are heavily shaped by cultural background, suggesting that there are variations in moral decision-making among individuals from different cultural back-

grounds (Liu et al., 2024; Vitolla et al., 2021). Awad et al. (2018) demonstrated that, by identifying distinct cultural clusters among respondents, these cultural differences are also present in moral decision-making within the MME dataset.

As significant proportions of the training corpora for LLMs originate from Western sources, it is expected that the moral compasses of LLMs align more with the cultural framework present in Western cultures (Takemoto, 2024; Ferrara, 2023). Moreover, previous research has found that the moral frameworks of LLMs often differ from non-Western moral frameworks (Liu et al., 2024; Vida et al., 2024).

To assess how differences in cultural frameworks affect ML model performance, which is the purpose of *SRQ3*, this research employs a dataset that contains only responses from non-Western participants. The assumption is that there are bigger differences between the moral decisions of humans and LLMs in this dataset compared to the dataset with all ethnicities, providing clearer patterns in the data. It is expected that the ML models are able to classify them more effectively, resulting in higher model scores. Based on this reasoning, current research hypothesizes that the models will perform better on the dataset exclusively containing data from non-Western respondents.

2.2 *Machine Learning Models (SRQ1)*

The adoption of LLMs is relatively recent, making the research focused on detecting their use - especially by using ML models - limited. Studies done so far have mainly focused on classifying LLM-generated texts versus human-written texts, rather than detection tasks more similar to the one in this research. Cingillioglu (2023) for example compared Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and a Neural Network (NN) when distinguishing written work made by LLMs and humans. They found that their NN achieved a highest accuracy of 0.933, compared to scores of 0.867 (LR), 0.9 (RF) and 0.927 (SVM). On similar tasks, Sanchez-Medina (2024) demonstrated that a RF model was able to achieve an accuracy of 0.841, while Kim and Desaire (2024) have achieved an accuracy of 93% with their XGBoost model.

As with LLM detection tasks, research employing ML models in the context of the MME is equally sparse. However, Agrawal et al. (2019) as well as Wiedeman et al. (2020) have utilized a NN, or so called Multilayer Perceptron (MLP), to predict moral decisions in the MME. Both have shown promising results, achieving accuracies of 0.771 (Agrawal et al., 2019) and between 0.72-0.81 (Wiedeman et al., 2020). In a comparable task, Singh, Murzello, Pokhrel, and Samuel (2024) have found that LR models struggle

to identify true positive instances, with F1-scores ranging from 0.33 to 0.61. Furthermore, Singh, Murzello, Lee, et al. (2024) have employed different ML models in order to classify utilitarian/non-utilitarian choices. They have found that a LR, SVM and simple NN all performed the same per situation, with macro F1-scores ranging from 0.73-0.96 across situations. Lastly, Steed and Williams (2020) employed a RF tasked with learning moral preferences in the MME dataset, achieving an accuracy of 69.8%.

This research focuses on employing LR, RF, SVM and a NN, as these have shown not only to be applicable in both LLM detection tasks and the MME context, but also show robust and promising results in those areas.

A recurring trend among prior studies is that the LR model underperforms compared to the other models, likely due to its simplicity and limited capabilities when presented with high dimensional data (Levy & O'Malley, 2020). In contrast, the NNs have shown superior performance in both the LLM as MME context. It is therefore hypothesized that this model will outperform the other models employed in this study (*SRQ1*).

As there has been no research done in trying to classify humans from LLMs in moral decision-making contexts, using ML models, it is quite difficult to set a true baseline performance of the models. The results presented in prior research, as described in this section, will therefore be taken as a loose guidance, and more emphasis will be placed on the models in this research beating the baseline model. This more fairly reflects the true performance of the models.

2.3 *Literature Gap*

While previous research has explored the capabilities of ML models in detecting LLM-generated, primarily written, content, a gap in literature investigating their effectiveness on non-textual moral data still exists. Specifically, no research investigated non-contextual representations of moral dilemmas with binary decisions, like those present in the MME. However, high-risk AI applications in the AV context often utilize structured data to generate binary SAVE/NO SAVE decisions, rather than detailed textual justifications. This highlights the prevalence of addressing this gap, offering new insights into ML models' capabilities in ethical decision-making contexts.

Additionally, while some research has explored the role of cultural influences in moral decision-making, none have examined cultural backgrounds as a factor in understanding differences between LLMs and humans in their moral frameworks. This study addresses this gap by investigating how cultural context affect model performance.

Summarizing, studies utilizing the MME have focused on predicting moral decisions using ML models but have not explored their ability in detecting LLM-generated responses from human ones. Current research addresses this gap, as illustrated in Figure 1, by examining this ability of ML models within the MME. Bridging this gap contributes to multiple fields by providing insights into ethical frameworks of LLMs, their potential differences from humans across cultures, and the suitability of ML models for detection tasks in ethical contexts.

3 METHOD

The methodology of this research has been summarized in a flowchart, depicted in Figure 2

3.1 *Dataset description*

3.1.1 *Moral Machine Dataset*

This research utilizes a subset of the SharedResponses.csv file, containing the dataset created by Awad et al. (2018) that was used in their original MME paper. This file is obtained from osf.com. The dataset encompasses moral decisions of participants from 233 countries, who were presented with scenarios where accidents resulting from sudden brake failure by an AV were unavoidable. In each scenario respondents had to choose between two outcomes, either saving pedestrians crossing the road or saving the passengers in the car. The characters and their attributes changed for every scenario. An example scenario is presented in Figure 3. This variability in scenario types enabled Awad et al. (2018) to research how attributes like gender, species or number of characters influenced the choice of participants of whom to save (Broska et al., 2024). The full dataset has approximately 70 million rows and 41 features, and a subset will be used.

3.1.2 *Large Language Model Moral Machine Dataset*

To research the capabilities of ML models in detecting the use of LLMs in moral decision-making, Takemoto (2024) 's data will be used. They repeated the setup of the original MME, but let various LLMs take the decisions instead of humans. Each LLM was presented with 50,000 scenarios in written form, and their decisions were collected in separate datasets – one for each LLM. The moral decisions from six LLMs were selected and combined into one dataset. The six datasets were collected from Takemoto's Dropbox.

Because of the controlled setup in which the data was collected in case of the LLMs, there are some discrepancies between the values in the MME and LLM dataset. Inconsistencies between them, like 'UserCountry3' only having 'Japan' as value in the LLM dataset, have been accounted for by removing these values in the original MME dataset, or by removing irrelevant features before modeling.

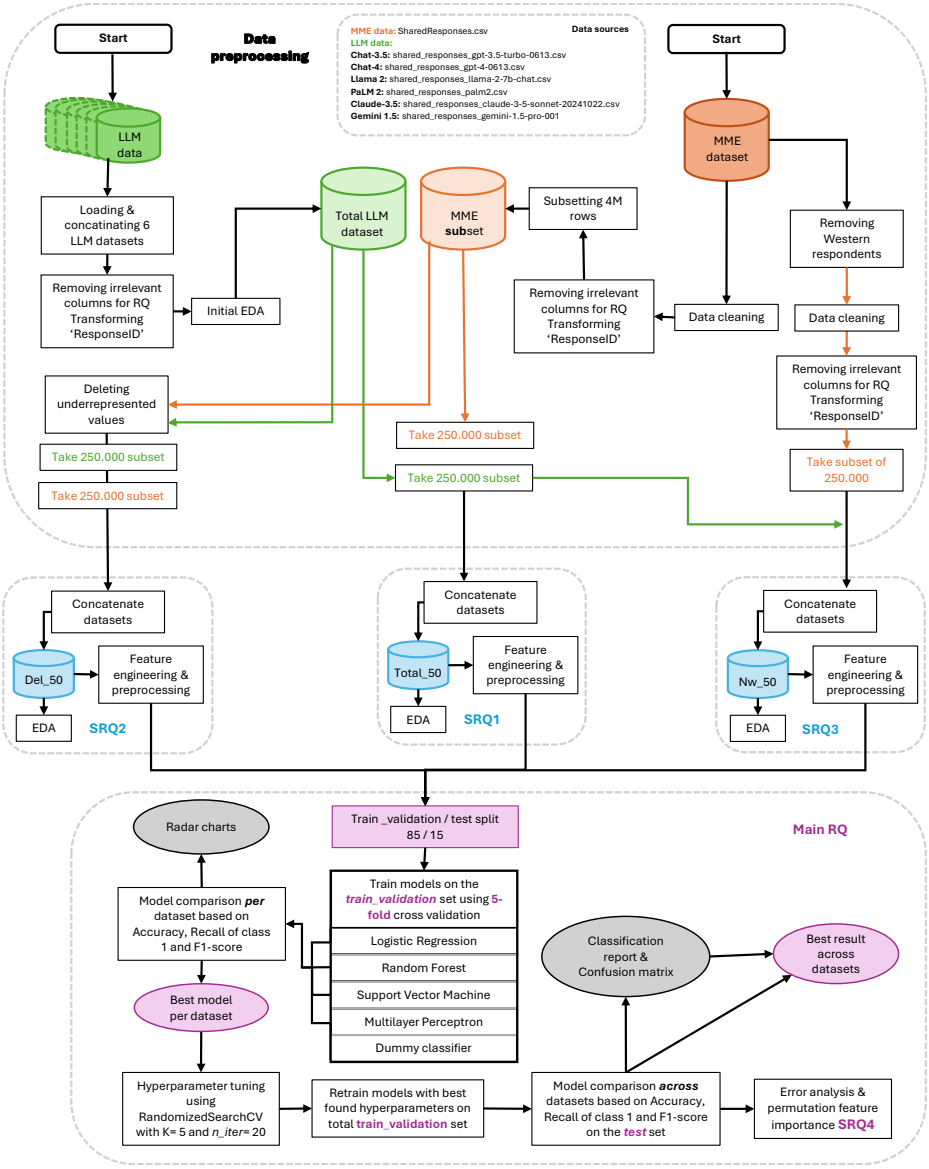


Figure 2: Flowchart summarizing research methodology

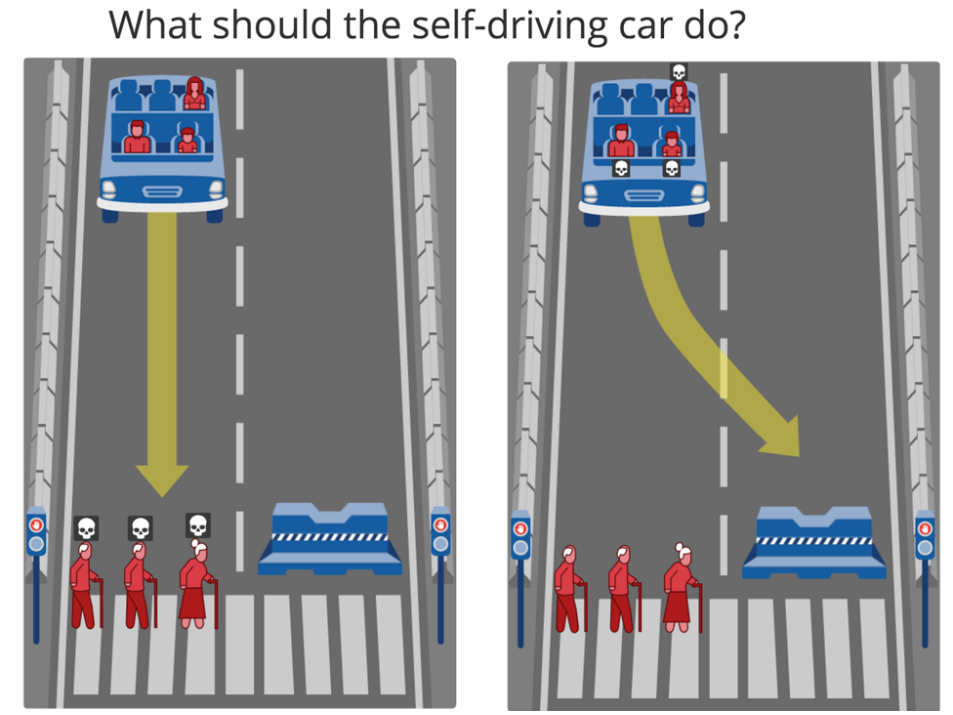


Figure 3: The Moral Machine interface with an example scenario. Source: Awad et al. (2018) (CC BY 4.0).

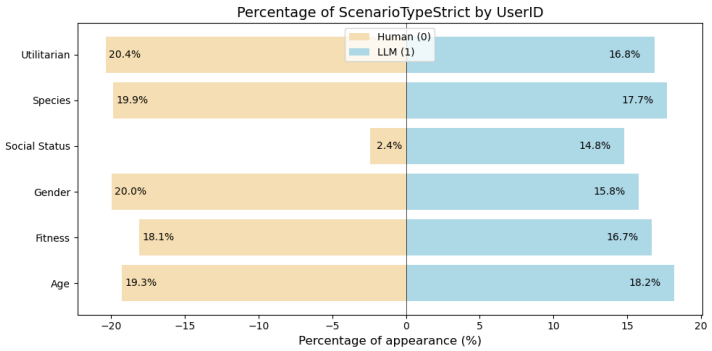
3.2 *EDA & Preprocessing*

3.2.1 *Missing values & outliers*

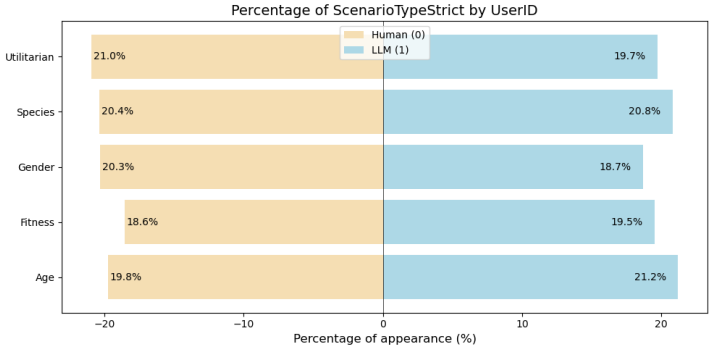
Missing value analysis revealed that only 0.02% of the data in relevant features were missing, and because of this insignificant size, they were all removed. Based on the summary statistics, no anomalies or outliers were detected.

3.2.2 *Distribution of features*

While exploring feature distributions per class in the target variable, significant differences between the occurrences certain feature values were found (see Figures 4 and 5). Due to this discrepancy of scenario occurrences, unrelated to their moral decision-making, there is a risk that the models may learn unintended associations. For example, the models could develop a tendency to predict 'LLM' whenever encountering the 'Social Status' attribute, as this value is considerably underrepresented in human responses. This bias would reflect the scenario composition instead of real patterns in moral decision-making. Attempts to create a dataset with identical scenarios for both human and LLM responses to isolate their



(a) Distribution of ScenarioTypeStrict *before* deleting underrepresented values



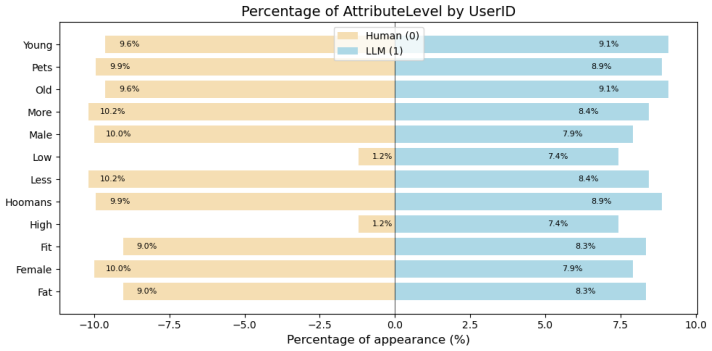
(b) Distribution of ScenarioTypeStrict *after* deleting underrepresented values (`df_deleted`)

Figure 4: Distribution of ScenarioTypeStrict

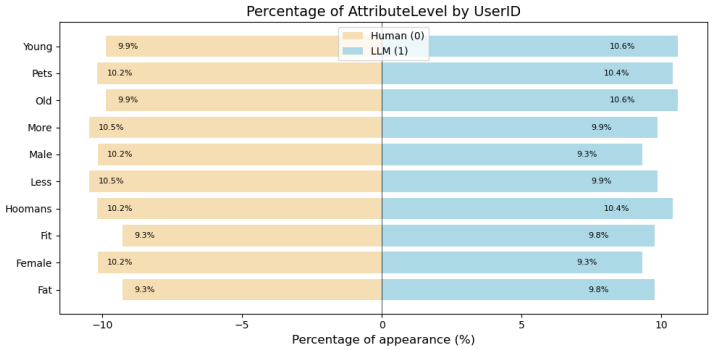
moral decision-making patterns were found to be computationally unfeasible. Therefore, to account for this potential source of error, a new dataset is created where instances with substantial distributional imbalances are excluded (`df_deleted`). The model performances will be compared to assess whether these discrepancies influence their learning, which is the aim of SRQ2.

3.2.3 *Multicollinearity*

Feature correlation heatmaps were used to assess multicollinearity among variables (Appendix B, page 46). Fu et al. (2020) state that Pearson's feature correlation values below 0.4 can be considered weak correlations, and typically don't require further action. In this research, only 0.6% of the feature pairs show a correlation greater than 0.3, and among these, the values remain below 0.6 (which Fu et al. (2020) consider moderate correlations). It was therefore decided not to address multicollinearity.



(a) Distribution of AttributeLevel *before* deleting underrepresented values



(b) Distribution of AttributeLevel *after* deleting underrepresented values `(df_deleted)`

Figure 5: Distribution of AttributeLevel

3.3 *Feature selection & Feature Engineering*

3.3.1 *Feature composition*

After removing features unnecessary for answering the main research question, the four constructed datasets (see Section 3.4) each consist of thirty features, prior to one-hot encoding. Of these, 28 describe the scenarios presented to respondents, one represents the moral choice the respondent made (`Saved`), and one reflects the respondent type - LLM or human (`UserID`). An overview of the 28 features describing the scenarios is provided in Appendix C (page 48).

To accomplish the goal of this research, which is evaluating the ability of ML models to detect whether a response to MME dilemma was generated by a human or LLM, the ML models are tasked with predicting the target variable `UserID` based on the scenario compositions and the moral choice. This binary target variable is encoded as 1 for all LLM responses, and 0 for all responses generated by humans.

3.3.2 *One-hot encoding & Normalization*

Two of the explanatory variables are categorical. As ML models aren't able to handle this type of data, one-hot encoding is used to transform the features into numerical representations. One-hot encoding is chosen because of its simplicity, interpretability and applicability in this context (Hussein et al., 2021; Almajid, 2022; Al-Shehari and Alsowail, 2021). It ensures that categorical variables are represented clearly without introducing ordinal bias (Lu, 2020). Moreover, an increase in dimensionality that is usually associated with one-hot encoding (Lu, 2020; Seger, 2018), is not of significant concern here, as the low cardinality of the categorical features results in only a small increase of the feature space. As all feature value fall within the range of 0-5 and no outliers were detected, normalization wasn't applied.

3.4 *Final datasets*

In order to answer the four SRQs, three datasets are constructed and compared. To ensure computational efficiency while simultaneously preserving predictive power of the models, each set contains 500.000 rows, accounting for 250.000 moral decisions. Table 2 summarizes their characteristics.

Dataset	Ethnicity	Underrepresented Values Deleted?
df_total	All	No
df_non-western	Only non-Western	No
df_deleted	All	Yes

Table 2: Details of ethnicity and scenario representation per dataset.

3.4.1 *df_total - SRQ1*

To answer SRQ1, a dataset consisting of participants from all ethnicities is constructed. Additionally, the target variable is completely balanced. The rows corresponding to LLM responses are randomly sampled from the complete LLM dataset, ensuring the scenarios remain intact, but without any modification or other selection criteria.

3.4.2 *df_non-western - SRQ3*

This balanced dataset contains only human respondents from a non-Western cultural background. The selection of 'non-Western' was based on a publication by the United Nations (United Nations, 2024), detailing which countries are considered 'Western'. By excluding respondents of the MME originating from those countries, it can be investigated to what extent culture influences model performance and what this means for the alignment of human and LLM ethical frameworks. The same subset of 250.000 LLM respondents as in the df_total dataset was used.

3.4.3 *df_deleted - SRQ2*

This dataset does not include the `Social Status` value in `ScenarioTypeStrict`, nor the 'High' and 'Low' values within the `AttributeLevel`. The same subset of 250.000 LLM respondents as in the df_total and df_non-western datasets was used. Removing those allows for assessing if discrepancies in scenario presentation influence the models learning (SRQ3). The distribution of these features is shown in Figures 4 and 5. As a result of this removal, this dataset contains three less columns after one-hot encoding.

3.4.4 *Class distribution*

The target variable is completely balanced across all three datasets, with both LLMs and humans represented by 250,000 rows. This even distribution prevents ML models from becoming biased towards one class, ensuring a clearer understanding of the models' true, unbiased performance.

3.5 *Modelling*

3.5.1 *Data partitioning*

Each dataset was divided into a training/validation set (85%) and a separate test set (15%). The training/validation set was used for model training and hyperparameter tuning, through a combination of k-fold cross-validation and random search. The final performance metrics are obtained by evaluating the best-performing models, after being retrained on the total training/validation set, on the test set.

3.5.2 *K-fold cross validation*

K-fold cross validation with a K of 5 is employed for all models on all datasets. K-fold cross validation is a commonly used technique in machine learning, used to obtain a robust estimation of the generalization error of classification models (Anguita et al., 2012). A 5-fold cross validation balances computational efficiency with robustness, providing reliable performance metrics that reflect a model's effectiveness on unseen data (Pinheiro et al., 2024, p. 8; Wong, 2015). Lastly, though a higher number of K, like 10 or 20, might generate more robust model performances, an increase in computational strain and run-time go paired with this, making them infeasible with the resource limitations present.

3.5.3 *Dummy Classifier*

To assess whether more complex models can learn additional interesting patterns in the data, a dummy classifier is used to set a baseline performance. The DummyClassifier from Scikit-Learn with an `Uniform` strategy is utilized for all datasets. The predictions are done uniformly at random - mimicking a random guesser - making it appropriate for balanced datasets.

3.5.4 *Model specifications*

The simple logistic regression (LR) will be employed using the default values of the parameters. These can be found here.

The Random Forest model will be implemented by using the RandomForestClassifier object of Scikit-Learn, with all parameters set to their default values. These can be found in the documentation of Scikit-Learn.

A Support Vector Machine (SVM) in case of a binary classification problem aims to find an optimal hyperplane that maximally separates the data-points by maximizing the margin between the two classes (Cortes, 1995). As this learning process is very computationally expensive, this research employs the ThunderSVM library created by Wen et al., 2018. By

exploiting GPUs, ThunderSVM enables more efficient processing without compromising the performance metrics or integrity of the model (Wen et al., 2018). The model will be initialized with all default parameters, meaning that Radial Basis Function (RBF) will be used as a kernel. RBF is suitable in this classification problem given the uncertainty of a linear decision boundary here (which RBF doesn't assume) (Van Belle & Lisboa, 2014). Further information on the parameters can be found in the ThunderSVM documentation.

A Multilayer Perceptron (MLP) is an Artificial Neural Network that consists of an input layer, one or more hidden layers and an output layer (Gardner & Dorling, 1998). It uses a backpropagation algorithm that aims to minimize the error between the predicted output and the ground truth (Rezaeian Zadeh et al., 2010). Gardner and Dorling (1998) furthermore state that MLP's can be trained to accurately generalize when presented with unseen data. This was demonstrated by Wiedeman et al. (2020) who achieved an accuracy of over 80% on the MME dataset. Agrawal et al. (2019) also developed a MLP model in a similar MME context, but the model proposed by Wiedeman et al. (2020) outperformed theirs. Furthermore, Wiedeman et al. (2020) conducted a more comprehensive hyperparameter search, contributing to their enhanced model performance. Moreover, Agrawal et al. (2019) did not specify their model specifications anywhere in their paper nor in an online code repository, making it infeasible to use as a guideline. Given these factors, this study replicates the model architecture of Wiedeman et al. (2020).

The MLP model used consists of an input layer, two dense layers - each with 64 neurons, ReLU activation and batch normalization - and an output layer with Sigmoid activation function. The details of the MLP model are described on page 51, in Appendix D, and is visually represented in Figure 6.

3.5.5 *Hyperparameter tuning*

Hyperparameter tuning is implemented to optimize ML model performance (Liao et al., 2022). Due to resource constraints, the hyperparameter search is limited to the best-performing model(s), identified through K-fold cross-validation. RandomizedSearchCV with a K=5 is used, as Random Search has a substantially lower computational cost, while being able to find as good or in some instances even better models compared to a Grid Search (Bergstra & Bengio, 2012). Details of the model and parameters to be tuned are specified in Section 4.2.

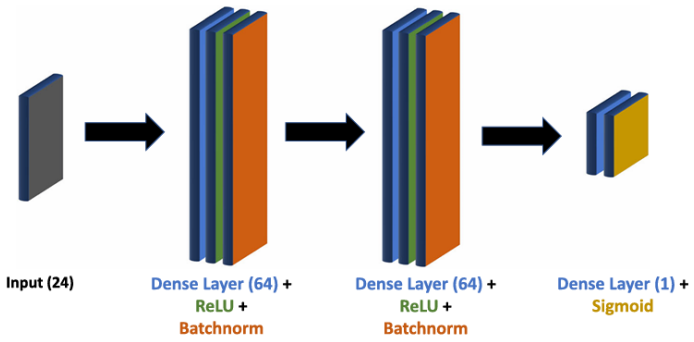


Figure 6: MLP model configuration employed in this research. Source: Wiedeman et al. (2020) (CC BY 4.0). *Note:* the input (24) was the size of the input vector in the Wiedeman paper, but here it is 45, reflecting the number of input features

3.5.6 *Performance evaluation*

When assessing model performances, the focus lies on the Recall for class 1 (LLMs), as the costs of false negatives (LLM classified as human) are high under current regulations. The macro F1-score is also evaluated to compare the models across datasets, as it accounts for class imbalance by giving equal weight to both classes, regardless of their frequency in the dataset. Accuracy is furthermore used to be able to compare model performances with prior research, as accuracy is most prevalent there. In addition, radar charts are used to visually compare the three metrics across the models, while confusion matrices help visualize error patterns on the test set.

To address SRQ4, permutation feature importance (PFI) will be employed on the models run on the test set. PFI evaluates feature importance by randomly shuffling feature values and observing the resulting decrease in model accuracy, indicating the feature's contribution to classification performance (Scikit-learn Developers, 2024). PFI is chosen over SHAP or LIME due to its straightforward implementation, scalability with larger datasets, and lower computational cost (Fumagalli et al., 2023). SHAP offers valuable insights into individual predictions, but due to its computational expense, this research employs alternative error analysis methods (see Section X) to achieve similar insights.

4 RESULTS

This chapter starts with an overview of results before hyperparameter tuning (Section 4.1), identifying the best-performing model per dataset for hyperparameter tuning. Next, the tuning process and rationale are outlined (Section 4.2), followed by a discussion on the results pre- and post-hyperparameter tuning per dataset (Sections 4.3 to 4.5). Then, a comprehensive comparison across models and datasets is presented (Section 4.6). Finally, the feature importance and further error analysis are presented (Section 4.7).

4.1 *Overview*

The scores presented are an average of the metric across all 5 folds. The standard deviation of the accuracy reflects the consistency of the results across folds. A low standard deviation suggests a stable accuracy good generalization. For tables 5, 7 and 9; *Recall* is that of class 1, *F1-score* is the macro F1-score, *Std Dev* is the standard deviation of accuracy and the Δ*baseline* is the change in Recall of class 1 (based on the unrounded metrics) compared to the dummy baseline.

Table 3 summarizes the the most important scores across all the models and datasets, before hyperparameter tuning.

Dataset	Baseline		LR		RF		SVM		MLP	
	Rec	F1	Rec	F1	Rec	F1	Rec	F1	Rec	F1
df_tot	0.498	0.502	0.477	0.591	0.678	0.704	0.477	0.644	0.572	0.651
df_nw	0.499	0.499	0.474	0.592	0.678	0.706	0.486	0.645	0.560	0.651
df_del	0.499	0.500	0.604	0.564	0.689	0.687	0.670	0.635	0.616	0.626

Table 3: Performance metrics across datasets for different models, pre-hyperparameter tuning.

Based on the overview above, the RF model outperforms all others across datasets, making it the chosen model for hyperparameter tuning. Detailed performance metrics are provided in following sections.

4.2 *Hyperparameter tuning*

The four RF hyperparameters tuned and their search spaces are detailed in Table 4.

To ensure a broad exploration of the hyperparameter space, a range of values rather than fixed values for each hyperparameter is specified. This reduces the risk of overlooking possibly effective intermediate values.

Given the limited application of RF models in the MME context, there are little to no references for selecting optimal hyperparameter ranges. Steed and Williams (2020) is one of the few that provides hyperparameter values of the RF model in the MME context, and as their model achieved an accuracy of almost 70%, it was ensured that current RF model parameters fall within their ranges

The number of trees are tuned to reduce the risk of overfitting while also balancing computational resources. The range is set from 50 to 300, as this is a balance between improved performance and the computation time that increases linearly with the number of trees (Probst et al., 2019).

The maximum depth of the tree controls how many layers the trees are allowed to grow to. Limiting the depth helps reduce overfitting and reduces the computation time significantly (Hidayat et al., 2019).

Decreasing the minimum number of samples required to split a node (`min_samples_split`) enables the trees to grow deeper, allowing it to learn more specific patterns. This might lead to overfitting and an increase in computation time however, which is why the range was increased to a maximum of ten.

Likewise, the minimum number of samples required to be at a leaf node (`min_samples_leaf`) is tuned to balance the risks of over- and underfitting, with the increase of computational cost that is added with the decrease in the number of leaves. An range of between one and five is typical.

Hyperparameter	**Values**
N_estimators	randint(50, 300)
Max_depth	[None] + list(randint(1, 30).rvs(10))
Min_samples_split	randint(2, 10)
Min_samples_leaf	randint(1, 5)

Table 4: Hyperparameters of the Random Forest model and their search space

The number of parameter settings that are sampled is set to 20, meaning that, K=5, a total of 100 fits are performed. The models with the best hyperparameter found during the random search are evaluated on the hold-out test to assess their final performance.

4.3 *df_total*

Table 5 shows the performance of the different models on the validation set of the df_total dataset, before hyperparameter tuning. The df_total dataset includes respondents of all ethnicities and retains underrepresented feature values.

	Evaluation Metric				
Model	**Recall**	**F1-score**	**Accuracy**	**Std Dev**	**Δbaseline**
Dummy	0.498	0.502	0.502	-	-
LR	0.477	0.591	0.597	0.001669	-4.18%
RF	**0.678**	**0.704**	**0.704**	**0.001068**	**36.09%**
SVM	0.477	0.644	0.655	0.000933	-4.31%
MLP	0.572	0.651	0.653	0.003174	14.92%

Table 5: Performance metrics for different models on the df_total dataset, pre-hyperparameter tuning.

As all model scores, except the Recall on class 1 of SVM and LR, exceed the baseline, it is implied that more complex models can learn additional useful patterns compared to random guessing. The standard deviations are all low, indicating consistent performance across validation splits.

The RF model outperformed all other models on all metrics, with a Recall score on class 1 of 0.678, a macro F1-score of 0.704 and an accuracy of 0.705. Figure 7 visually represents this in a radar chart, showing that the polygon of the RF model covers more grid space than the polygons of the others. Moreover, the Recall of the RF model surpasses the baseline with 36.09%, which is more than 20 percent points higher than the next best-performing model, the MLP. Thus, the RF model is the best choice in this dataset.

The SVM and MLP models perform similarly in terms of F1-score and accuracy, but differ significantly when considering the Recall for class 1 (0.477 vs. 0.572 respectively). This suggests that the MLP model is more effective at detecting LLM responses, as it captures more true positives in class 1. Additionally, the MLP model balances both classes better, whereas the SVM model tends to focus on identifying human responses (class 0) more effectively. Lastly, the LR model shows the poorest performance in comparison to the others.

4.3.1 *Hyperparameters and test set performance df_total*

Table 6 summarizes the found hyperparameters of the RF model on the df_total dataset.

The found hyperparameters allow the RF model to capture specific, fine-grained patterns in the data. The high number of *N_estimators* (293) suggests that the model benefits from reduced variance, while a *maximum depth* of 26 balances increasing model complexity with overfitting. Additionally, the *min_samples_split* (2) and *min_samples_leaf* (1) values indicate that the model performs best when allowed to grow deep trees with single-sample leaf nodes to learn subtle patterns when training.

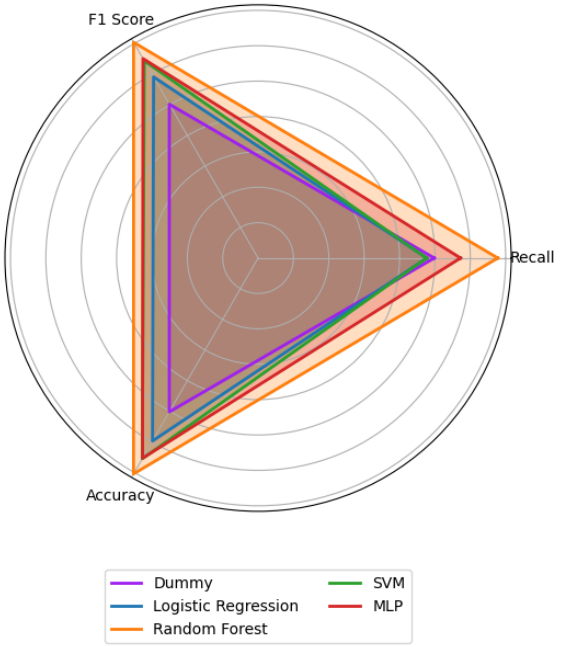


Figure 7: Radar chart comparing all the models on the df_total dataset

Hyperparameter	Values
N_estimators	293
Max_depth	26
Min_samples_split	2
Min_samples_leaf	1

Table 6: Best hyperparameters found on the df_total dataset

This configuration that allows for capturing complexity also increases the risk of overfitting. Comparing training and test results it becomes evident that this is also the case here (see also Figures 13 and 14). The F1-score dropped from 0.778 (training) to 0.711 (test), and the Recall for class 1 decreased from 0.745 to 0.673. These differences are relatively small however, indicating that the generalization power was not greatly affected. Moreover, the final test set metrics are close to the pre-tuned validation scores, suggesting that overfitting, while present, is not of major concern here. This also indicates that, overall, hyperparameter tuning did not significantly improve model performance.

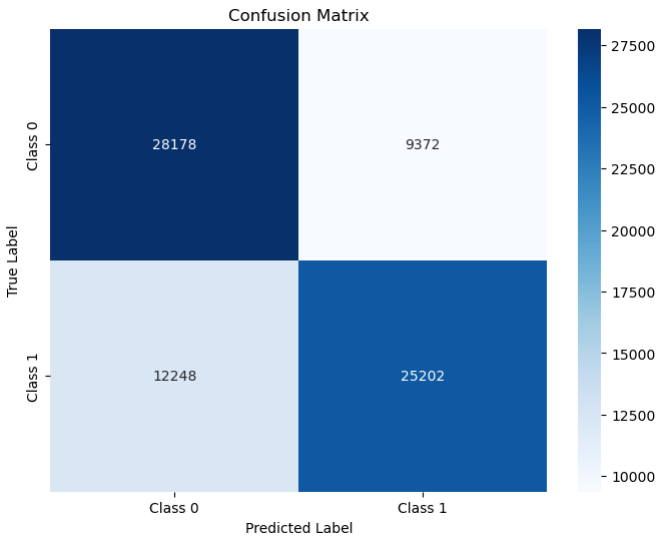


Figure 8: Confusion matrix of the tuned RF on the test set of df_total

4.4 *df_non-western*

The predictive performances of each model on the validation set of the non-Western dataset are presented in table 7. This dataset includes respondents of only non-Western backgrounds, but still retains underrepresented feature values.

	Evaluation Metric				
Model	**Recall**	**F1-score**	**Accuracy**	**Std Dev**	**Δbaseline**
Dummy	0.499	0.499	0.499	-	-
LR	0.474	0.592	0.598	0.001387	-5.11%
RF	**0.678**	**0.706**	**0.707**	**0.001181**	**35.72%**
SVM	0.486	0.645	0.656	0.001508	-2.65%
MLP	0.560	0.651	0.654	0.002192	12.11%

Table 7: Performance metrics for different models on the df_non-western dataset, pre-hyperparameter tuning.

As with the df_total dataset, all models, except the Recall on class 1 of SVM and LR, surpassed the performance of the baseline model, with the RF model achieving the highest performance across all metrics. RF demonstrated superior Recall of class 1 (0.678), macro F1-score (0.706) and accuracy (0.707), with which it outperforms the other models. Moreover, its 35.72% Recall increase over the baseline, which is significantly larger than the others, highlights the RF model's ability to distinguish human

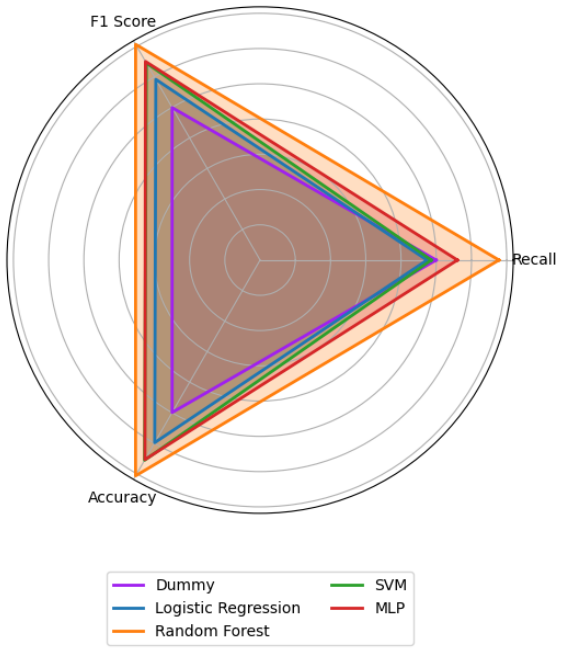


Figure 9: Radar chart comparing all the models on the df_non-western dataset

from LLM responses in a dataset exclusively featuring non-Western human respondents.

The SVM and MLP models show comparable performance on F1-score and Accuracy. The MLP however achieves a higher Recall score (0.560) than the SVM (0.486), suggesting that the MLP model is better in classifying LLMs compared to the SVM model.

The low standard deviations for all models indicate consistent and reliable performance across folds. These results (see also Figure 9), highlight RF as the best-performing model, followed by MLP and SVM.

4.4.1 *Hyperparameters and test set performance df_non-western*

The best hyperparameters are presented in table 8.

Hyperparameter	**Values**
N_estimators	248
Max_depth	26
Min_samples_split	2
Min_samples_leaf	1

Table 8: Best hyperparameters found on the df_non-western dataset

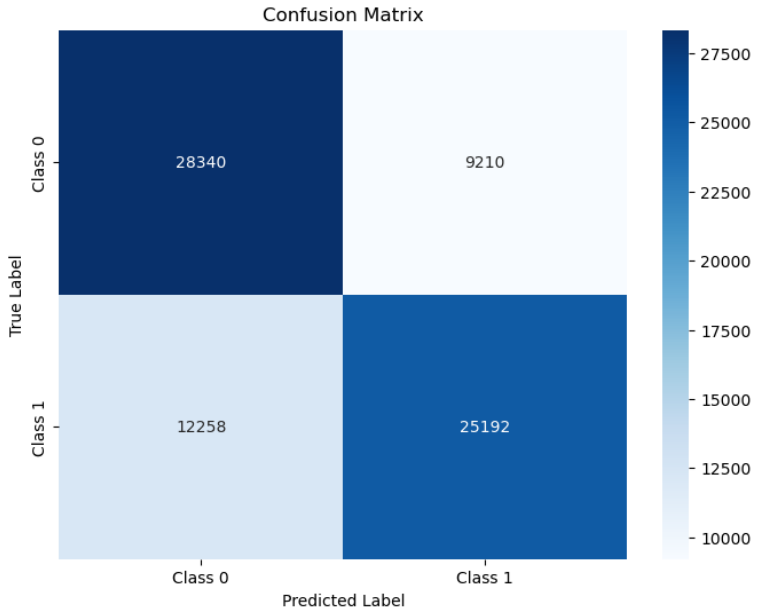


Figure 10: Confusion matrix of the tuned RF on the test set of df_non-western

What becomes evident is that the these parameters are very similar to those considered optimal in the df_total dataset, with the only difference being the *N_estimators* being somewhat lower (248 vs 293). Though lower, 248 trees is still considered to be on the high end. This implies that the model in this dataset also benefits from the ability to capture more complex patterns in the data, albeit a bit more contained because of this lower number of trees allowed.

Where the Recall test score is somewhat lower than in the pre-tuned model (0.673 vs. 0.678), both the F1-score and accuracy on the hold-out test set are slightly higher (see also Table **??**). This suggests that the model, when retrained with optimal hyperparameters, is a little less effective at classifying class 1 (LLMs), but slightly better at classifying class 0 (human responses) (see also Figure 10).

As illustrated in Figures 13 and 14, these model hyperparameters resulted in slight overfitting on the training set compared to the test set. Nonetheless, these differences in performance are minor and are therefore not considered to be of major concern.

4.5 *df_deleted*

Table 9 summarizes model performances on the deleted_50 dataset, pre-hyperparameter tuning. This dataset includes respondents from all cultural backgrounds, but excludes underrepresented feature values.

	Evaluation Metric				
Model	**Recall**	**F1-score**	**Accuracy**	**Std Dev**	**Δbaseline**
Dummy	0.499	0.500	0.500	-	-
LR	0.604	0.564	0.564	0.001117	20.90%
RF	**0.689**	**0.687**	**0.687**	**0.001400**	**38.05%**
SVM	0.670	0.635	0.636	0.000859	34.12%
MLP	0.616	0.626	0.626	0.002217	23.33%

Table 9: Performance metrics for different models on the df_deleted dataset, pre-hyperparameter tuning.

The performance metrics of the models on the deleted_50 dataset conform to the general trend, with the RF model outperforming all others and the standard deviations being low. The RF model has a Recall score on class 1 of 0.689, and a macro F1-score and accuracy of both 0.687, as also presented in Figure 11. Where in the other models the LR and SVM did not surpass the Recall baseline, they both do here. This suggests that the models in this dataset are better at detecting LLMs than a random guessing baseline.

Interestingly, the Recall scores of the LR, RF and SVM models is higher than their accuracy and F1-score, indicating that they are better at classifying LLM responses compared to humans. This deviates from most results seen so far, as for all models on both datasets the opposite was true. It suggests that deleting underrepresented values positively influences the correct classification of LLMs compared to humans respondents in the LR, RF and SVM models.

Moreover, the metrics per model are quite similar, implying a balanced learning process in this dataset. They are not favoring one class over the other, which would have been reflected in the Recall for class 1 being much higher or lower than the accuracy and F1-scores. Despite the lower F1-score and accuracy of the SVM model, its Recall score of 0.670 is quite close to that of the RF model (0.689). This shows that the SVM is performing competitively with the RF model in detecting class 1 (the LLM responses), more so than in the other datasets.

4.5.1 *Hyperparameters and test set performance df_deleted*

The optimal hyperparameters found on the df_deleted dataset are shown in table 10

Notably, the *N_estimators* parameter is quite low compared to the other datasets (93 compared to 248 or higher). Also the *min_samples_leaf* is being quite high in comparison to the other datasets (7 vs. 1). This suggests that the model apparently benefits from regularization in the form of limiting

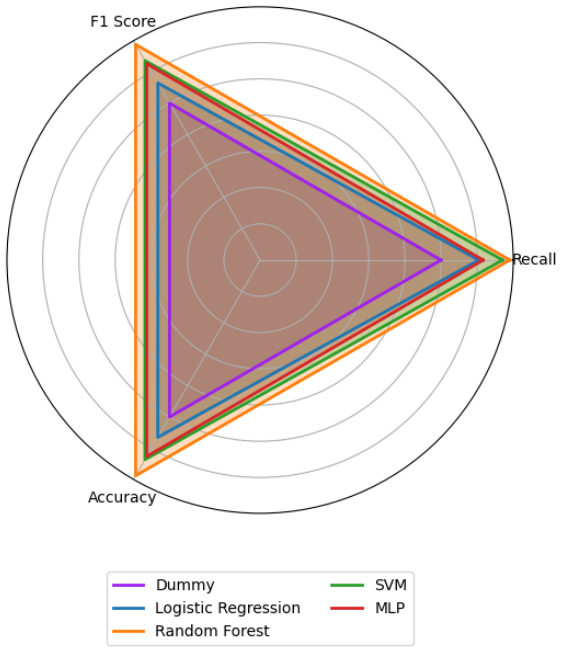


Figure 11: Radar chart comparing all the models on the df_deleted dataset

Hyperparameter	Values
N_estimators	93
Max_depth	None
Min_samples_split	2
Min_samples_leaf	7

Table 10: Best hyperparameters found on the df_deleted dataset

the number of trees and indicating the minimum number of samples in each leaf node. The values of *max_depth* and *min_samples_split* being 'None' and 2 respectively indicate that the model is being regularized through the other two parameters, while still benefiting from capturing nuanced patterns through these other parameters.

Although some of the hyperparameters apply a form of regularization, the retrained model still slightly overfits on the hold out test set (see Figures 13 and 14). These differences are minimal and don't raise substantial concerns here.

As was the case for the pre-tuned model on this dataset, the scores of the different metrics on the test set all lie closely together. This suggests that the model doesn't favor one class over the other, indicating a balanced learning approach of the RF model, as shown in the confusion matrix and error patterns in Figures 12 and 15c.

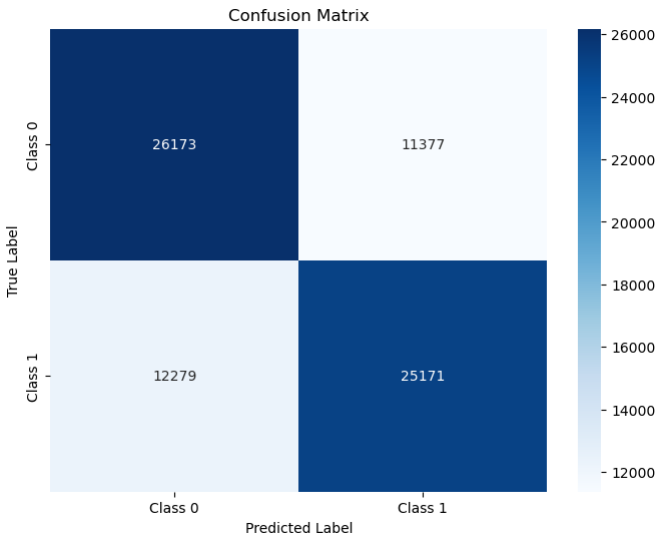


Figure 12: Confusion matrix of the tuned RF on the test set of df_deleted

4.6 *Comparison across datasets*

Dataset	Recall	F1 Score	Accuracy
df_total	0.673	0.711	0.712
df_non-western	0.673	0.713	0.714
df_deleted	0.672	0.685	0.685

Table 11: Performance Metrics for the test sets for all three datasets

When comparing the Recall and F1-scores across all models and datasets, on both the validation sets and test sets, presented in tables 5 and 11, a few things emerge.

Firstly, the Recall scores in the deleted_50 dataset are notably higher than in the other datasets. This implies that deleting the underrepresented values enabled the models to accurately detect more LLM responses. It is important to note that though the Recall scores have improved, the F1-scores in this dataset are slightly lower compared to the other datasets, implying a trade-off between Precision and Recall.

Secondly, the deleted_50 dataset shows more consistent Recall scores (range 0.085 versus 0.201 and 0.204 in other datasets), as models like LR, SVM, and MLP perform better here compared to the others. This improvement suggests the deleted values introduced variability in the other

datasets that hindered these models' ability to detect LLM responses.

Thirdly the RF has a closer alignment between F1 and Recall scores across all datasets than other models. This suggests it is able to balance the classes in various settings. Moreover, the RF model performs similarly across the datasets, both on the validation sets as on the hold-out test set. This is unlike the other models, as they show more variability depending on the dataset they are run on.

Fourth, the scores on the test set are very similar to the scores on obtained from the cross validation ran on the pre-tuned models. This signifies that the RF models, even in different datasets, hold a high generalization power.

Fifth, while the Recall scores are similar, the F1 and accuracy scores of the df_deleted dataset are lower than that of the other datasets. This suggesting that the RF model in the df_deleted dataset is equally capable in detecting responses of LLMs, but lacks in its ability to classify human responses. This becomes also evident in the confusion matrix (Figure 12), where the False Negatives are higher than in those of the other datasets.

Lastly, the observed overfitting may be due to the models being trained on the combined train_val dataset, increasing the training set from 70% to 85% of the total data. This could explain the higher training scores, as larger training datasets often enhance model performance by enabling it to capture more information (Junqué de Fortuny et al., 2013).

4.7 *Error analysis and feature importance (SRQ4)*

Figure 15 shows the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) for each `ScenarioTypeStrict` category on the test set, along with the difference between the total Trues and Negatives in percentage points (p.p). A higher p.p. difference suggests bigger divergence between human and LLM moral frameworks, as apparently the RF model can better distinguish between them. Several key things are noted from these error representations.

The high TP count in the Social Status category for the df_total and df_nw datasets suggests that the RF model does in fact learn unintended associations from scenario inconsistencies, thereby significantly boosting the Recall score for class 1 in these datasets.

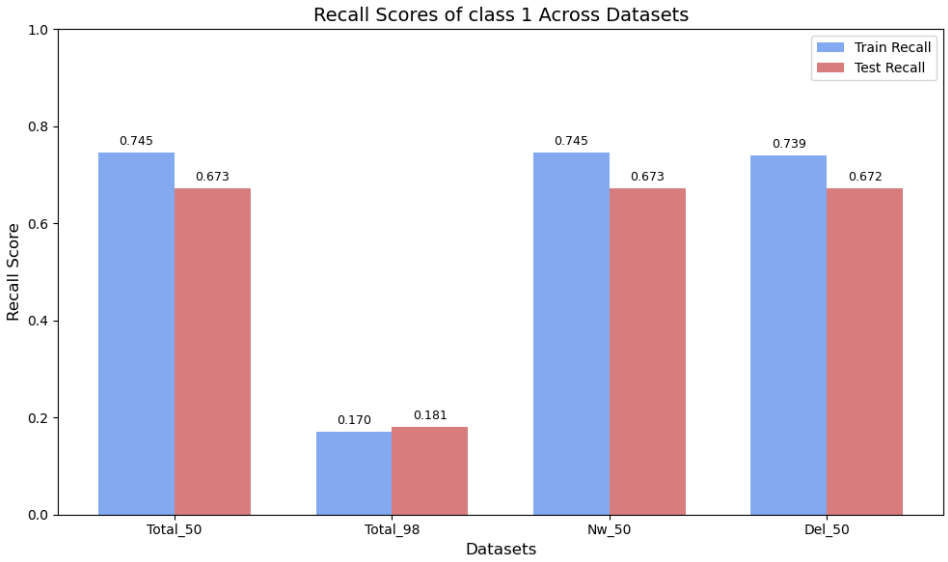


Figure 13: Comparison of class 1 Recall score of the best RF model found during hyperparameter tuning on the train and test set, for all the datasets

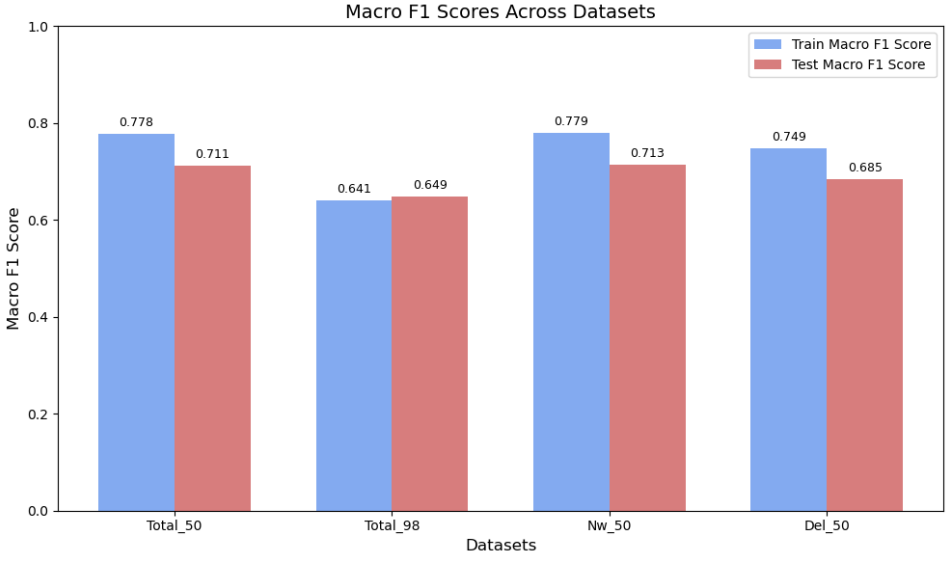


Figure 14: Comparison of Macro F1-score of the best RF model found during hyperparameter tuning on the train and test set, of all the datasets

Another key observation is the apparent difference in moral frameworks between LLMs and humans regarding utilitarianism. Across all graphs, a significant difference exists between correctly and incorrectly identified users in this category (69.4–75.7 p.p.) compared to other categories, highlighting how human moral decisions differ from those of LLMs in such scenarios. This importance is reinforced by the permutation feature importance (Figures 16, 17 and 18), where the relevance of features like `NumberOfCharacters` and `DiffNumberOfCharacters` reflect the Utilitarian scenario type.

Moreover, there is little to no difference in error patterns between the df_total and df_nw datasets. This indicates that, at least for the scenarios presented here, the moral frameworks of humans from different cultural backgrounds are not substantially different — at least not substantial enough for the RF model to detect.

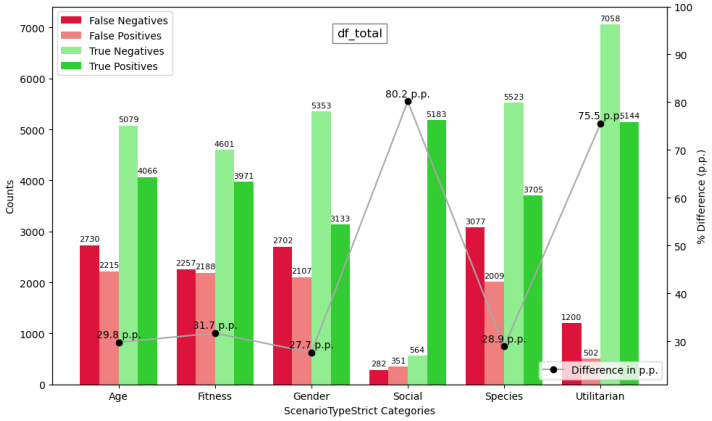
In the df_del dataset (see Figure 15c) the TNs and TPs lie closer together suggesting a more balanced classification of humans and LLMs in comparison with the other datasets. Lastly, in some categories in this df_del dataset, TPs exceed TNs, indicating the RF model is slightly better at identifying LLMs in these cases.

4.7.1 *Feature Importance*

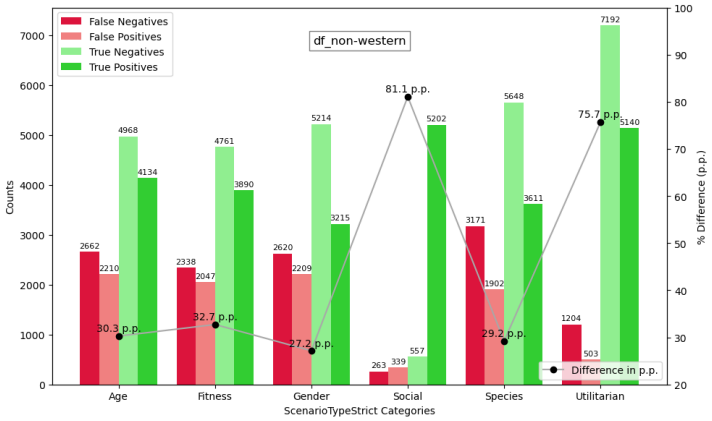
Figures 16, 17 and 18 show the feature importance on the test set of the three datasets.

Similar features are found to be important across all datasets, with `NumberOfCharacters` and `DiffNumberOfCharacters` are the first and third most important with similar importance scores among datasets. This implies that, as mentioned above, LLM and human ethical frameworks differ to a certain extent in utilitarian type scenarios. The feature `Saved` is the second most significant in all datasets, suggesting that the RF model uses users' moral decisions (whom to save) as a proxy for underlying ethical patterns. Its importance is higher in the df_deleted dataset (0.0546) compared to the df_nw (0.0489) and df_total (0.0473) datasets, indicating it relies more on this feature in the df_deleted dataset.

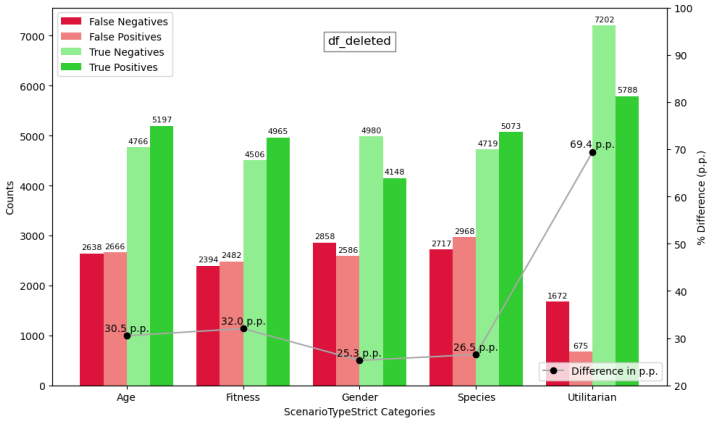
The features PedPed, Intervention, and CrossingSignal exhibit moderate importance, with their scores across all datasets ranging from 0.0330 to 0.0401. These values are noticeably lower compared to the most important feature, which has scores exceeding 0.0700. The remaining features demonstrate significantly lower importance scores, suggesting they contribute less to the model's predictions.



(a) Distribution of True and False classifications across ScenarioTypeStrict categories in the df_total dataset



(b) Distribution of True and False classifications across ScenarioTypeStrict categories in the df_non-Western dataset.



(c) Distribution of True and False classifications across ScenarioTypeStrict categories in the df_deleted dataset

Figure 15: Distribution of True and False classifications in the ScenarioTypeStrict feature across datasets, with percentage point (p.p) differences between them. *Note:* 'Social' reflects the 'Social Status' category

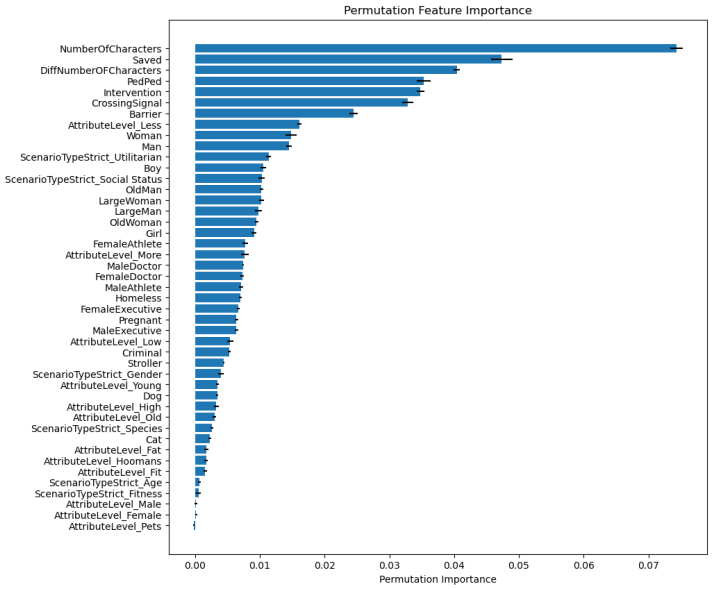


Figure 16: Permutation feature importance on the df_total test set.

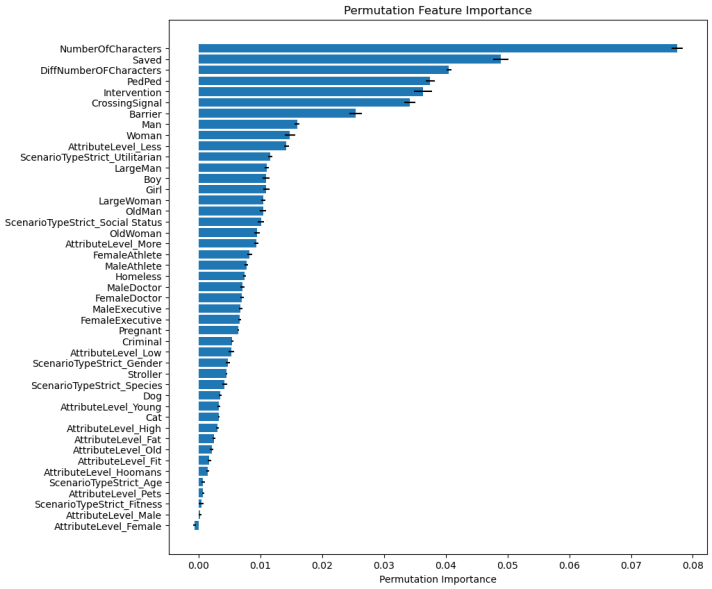


Figure 17: Permutation feature importance on the df_nw test set.

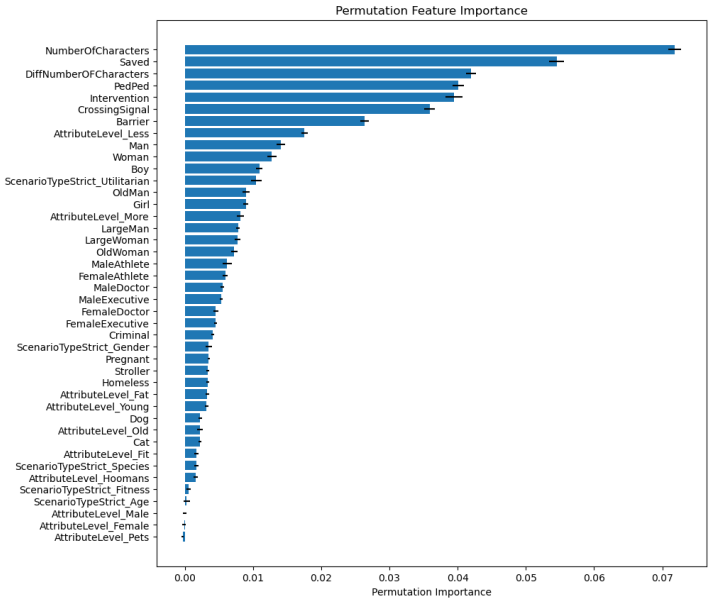


Figure 18: Permutation feature importance on the df_deleted test set.

5 DISCUSSION

This research aimed to assess the extent to which Machine Learning (ML) models can accurately detect whether responses to Moral Machine Experiment (MME) dilemmas were generated by humans or Large Language Models (LLMs), across different contexts. To this end, performance metrics of four models on three datasets were compared. Additionally, to gain more insight into where moral frameworks of humans and LLMs differ, an error analysis accompanied with feature importance analysis was employed.

5.1 *Answering research questions*

SRQ1 - It was anticipated that the MLP model would outperform the other models, as it had shown that superior performance in both the LLM and MME context. However, the RF model was the model that outperformed the LR, SVM and MLP across all datasets. It achieved accuracies between 0.685-0.714 and Recall scores for class 1 ranged from 0.672-0.673. These results align with Steed and Williams (2020), who reported similar RF accuracy of almost 70% in the MME context. The RF's robustness and adaptability can be attributed to its reliance on ensembles of decision trees, which inherently reduces the impact of noisy features (Biau, 2010). This likely also explains why the RF, compared to other models, was

less affected by the removal of underrepresented values in the df_deleted dataset.

Moreover, the LR model consistently performed the worst, as expected. This aligns with prior findings, where the LR also underperformed compared to more complex alternatives (Singh, Murzello, Pokhrel, and Samuel, 2024; Cingillioglu, 2023). As the model configuration created by Wiedeman et al. (2020) was copied, this study expected the accuracies of the MLP models on the different datasets to be similar to those found in their research (0.72-0.81). However, range of the model's accuracy for the MLP in this research lie between 0.626 and 0.654, which might be attributed to the difference in classification tasks. Lastly, while the SVM model had shown potential in earlier research (Singh, Murzello, Lee, et al., 2024), with accuracy scores up to 0.96, it did not live up to these expectations. This could be explained by the fact that this research does not predict moral decisions likeSingh, Murzello, Lee, et al. (2024) did, but rather tries to identify by whom the moral decisions were made.

SRQ2 - It was hypothesized that models trained on datasets with underrepresented values (df_total and df_nw) might learn unintended associations, leading to better performance due to reliance on shortcuts, and that removing these values (df_deleted) would harm performance. This was partially supported, as the F1-score and accuracy for df_deleted (both 0.685) were slightly lower than for the other datasets (0.711–0.714). However, the Recall score for class 1 remained consistent (0.672 vs. 0.673). Error analysis and confusion matrices showed that while the RF model classified LLMs better in df_deleted, its ability to classify human responses (class 0) slightly declined, explaining the similar Recall but lower F1 and accuracy scores to the other datasets. This suggests the removed values introduced noise rather than useful information and that removing them did not significantly harm the models effectiveness but instead highlighted its adaptability to learn without relying on shortcuts.

SRQ3 - Based on research done by Vida et al. (2024) and Liu et al. (2024), which stated that LLMs differ in their moral frameworks from non-Westerners, this research hypothesized that ML models would be able to classify LLMs and human responses more effectively in a dataset containing only non-Western respondents compared to one with respondents from all backgrounds. However, this was not supported by the analysis presented above, as model performance was nearly identical across the two datasets. Both achieved Recall scores of 0.673, and F1-scores and accuracies only differed 0.002 between them. These findings indicate that the six LLMs used in this study exhibit less moral bias toward a Western moral framework than anticipated. This underscores that while differences between LLM and human decision-making exist - demonstrated by models achieving scores

well above random guessing – these differences appear less influenced by respondents' cultural backgrounds than initially expected.

SRQ4 - They might be more related to the moral scenario types, as previous research has shown that human and LLM moral frameworks differ in certain moral dilemmas, particularly in utilitarian, fitness, and species scenario types (Takemoto, 2024; Vida et al., 2024). It was therefore expected that error analysis and feature importance would highlight these scenarios as key in RF model classifications across the three datasets. Results confirmed this for the utilitarian scenario type, which showed the largest difference between correctly and incorrectly classified instances across all datasets (over 69.4 percentage points, compared to a maximum of 32.7 in other categories). Utilitarian features were also consistently found to be the most important across datasets, underscoring this as an area where human and LLM moral frameworks still differ. However, no significant differences were found for fitness or species scenarios, suggesting the LLM data used in this research aligns more closely with human preferences in these moral dilemma types.

5.2 *Societal implications*

For policy makers, this research offers actionable insights by presenting ML models as possible tools for detecting the unreported use of LLMs to make moral decisions in high-risk AI areas, especially in the AV industry. Although the performances of these models are not flawless, in balanced situations they are able to show significant detection capabilities, with Recall scores of at least 67.2%. This shows that they are able to offer additional detection capabilities in various contexts, compared to the random guessing that might be employed now. Highlighting areas where LLMs and humans still differ in their moral frameworks can help encourage LLM development to align more with humans values. Finally, by, to some extent, facilitating the identification of unreported LLM usage, this research contributes to the improvement of safety in high-risk AI contexts.

5.3 *Scientific implications*

This research contributes to scientific literature by improving understanding of how four established ML models behave when asked to detect responses generated by LLMs within the context of the MME. By pioneering in this direction, current research bridges the gap between the fields of LLM research, moral decision-making frameworks, and ML detection techniques, making it a novel contribution to existing literature. Moreover, by comparing the models in differently configured datasets, this research

provides insight into how they perform under different conditions. Lastly, the findings enhance understanding of where the moral frameworks of humans and LLMs might differ or align, indicating where further refinement is needed to align them. This work lays a foundation of future research surrounding detection of unreported LLM use in high-risk AI domains.

5.4 *Limitations and future research*

Though this research has both useful scientific and societal contributions, there are some limitations that need to be addressed.

Firstly, only data from 5 different LLMs was utilized in this research. However, there are currently dozens of LLMs out there (Guinness, 2024), who likely have different moral preferences from each other and from humans (Takemoto, 2024; Vida et al., 2024; Liu et al., 2024; Ahmad and Takemoto, 2024). Prospective studies could take data from other and/or more recent versions of LLMs, and assess both how other models hold themselves against moral frameworks of humans, and how the moral stances of these models might change over time.

Moreover, the scope of this research was restrained by computational resources, limiting the ability to explore more advanced models and techniques. More advanced ML models and analyses, like XGBoost (where Singh, Murzello, Pokhrel, and Samuel (2024) has found potential in in similar contexts), deep Neural Networks or other non-ML approaches, might improve model scores. Future researchers can take this upon themselves to test this. Additionally, techniques that optimize current model scores can be further investigated, as current research was limited to randomized search.

Furthermore, while this study did perform an error analysis that included feature importance, it focused mainly on the `ScenarioTypeStrict` feature, as prior research (Takemoto, 2024; Vida et al., 2024) highlighted differences between humans and LLMs in certain categories of this feature. However, exploring other scenario features that reflect moral preferences, like `CrossingSignal` that indicates the value respondents put on saving characters that (il)legally cross the road, might reveal additional differences in ethical decision-making between humans and LLMs. Subsequent studies might also achieve similar insights by repeating current study, but ensuring that both LLMs and humans are presented with identical scenarios. This could isolate their moral differences, highlighting discrepancies between their ethical frameworks.

Lastly, this research limited itself to the AV industry, but exploring these models in other high-risk AI sectors like banking or healthcare could be valuable for future research.

6 CONCLUSION

The main goal of this research was investigating how effectively various machine learning (ML) models can detect whether a response to a Moral Machine Experiment (MME) scenario was generated by a Large Language Model (LLM) or by a human, in different contexts. It was found that, among Logistic Regression, Random Forest (RF), Support Vector Machine, and Multilayer Perceptron models, the RF model was the most effective in this task with macro F1-scores ranging from 0.685 to 0.713 across three different datasets. These datasets included respondents from all ethnicities (df_total) and only non-Western respondents (df_nw), reflecting hypotheses risen from current literature. The third had underrepresented values removed (df_deleted), to account for potential learning biases related to these underrepresented values.

Analysis suggests minimal differences in moral frameworks across cultural backgrounds, as error patterns and model scores of the df_total and df_nw datasets were similar. Moreover, the RF model demonstrated the ability to partially mitigate learning biases present in other datasets. This is reflected in its slightly lower F1-score (0.685 versus around 7.12), indicating greater difficulty in classifying humans. However, the RF achieved a similar Recall score (0.672 versus 0.673), reflecting its robustness and adaptability in detecting LLMs. Lastly, features representing utilitarian moral scenarios were found to be the most important across all datasets, highlighting these dilemmas a key area of divergence between human and LLM moral frameworks.

Overall, this research demonstrates the extent to which different machine learning models can be utilized to detect the use of LLMs in high-risk AI applications, such that of moral decision-making in AV's. Aside from adding to current literature by revealing ML detection capabilities in moral decision contexts and the differences between moral frameworks of humans and LLMs, this study responds to calls for stricter regulatory oversight in high-risk AI industries. Ultimately, this work contributes to a safer society by, to some extent, enabling the detection of the unreported LLM use in moral decision-making processes.

REFERENCES

African Union. (2024, July). Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity [Accessed: 2024-09-09]. https://au.int/sites/default/files/documents/44004-doc-EN-_Continental_AI_Strategy_July_2024.pdf

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*.

Ahmad, M. S. Z. b., & Takemoto, K. (2024). Large-scale moral machine experiment on large language models. *arXiv preprint arXiv:2411.06790*.

Almajid, A. (2022). Multilayer perceptron optimization on imbalanced data using svm-smote and one-hot encoding for credit card default prediction. *Journal of Advances in Information Systems and Technology*, *3*(2), 67–74. https://doi.org/10.15294/jaist.v3i2.57061

Al-Shehari, T., & Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, *23*(10). https://doi.org/10.3390/e23101258

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The 'k' in k-fold cross validation. *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 77–82. http://www.i6doc.com/en/livre/?GCOI=28001100967420

Association of Southeast Asian Nations. (2024). ASEAN Guide on AI Governance and Ethics [Accessed: 2024-09-09]. https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf

Awad, E., Dsouza, S., & Kim, R. (2018). The moral machine experiment. *Nature*, *563*, 59–64. https://doi.org/10.1038/s41586-018-0637-6

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

Broska, D., Howes, M., & van Loon, A. (2024, August). The mixed subjects design: Treating large language models as (potentially) informative observations [Preprint on OSF]. https://doi.org/10.31235/osf.io/j3bnt

Cingillioglu, I. (2023). Detecting ai-generated essays: The chatgpt challenge. *The International Journal of Information and Learning Technology*, *40*(3), 259–268.

Cortes, C. (1995). Support-vector networks. *Machine Learning*.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning

techniques. *Journal of Big Data*, *2*(1), 23. https://doi.org/10.1186/s40537-015-0029-9

European Parliament. (2023, June). EU AI Act: First Regulation on Artificial Intelligence [Accessed: 2024-09-09]. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

European Parliament. (2024, March). Artificial Intelligence Act: MEPs adopt landmark law [Accessed: 2024-09-09]. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

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

Fu, T., Tang, X., Cai, Z., Zuo, Y., Tang, Y., & Zhao, X. (2020). Correlation research of phase angle variation and coating performance by means of pearson's correlation coefficient. *Progress in Organic Coatings*, *139*, 105459. https://doi.org/10.1016/j.porgcoat.2019.105459

Fumagalli, F., Muschalik, M., Hüllermeier, E., & Hammer, B. (2023). Incremental permutation feature importance (ipfi): Towards online explanations on data streams. *Machine Learning*, *112*(12), 4863–4903.

Gardner, M. W., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627–2636.

Guinness, H. (2024). What is multimodal ai? large multimodal models, explained [Accessed: 2024-11-30, published on 2024-08-05]. https://zapier.com/blog/best-llm/

Hidayat, S., Ashari, A., & Putra, A. E. (2019). Depth limitation and splitting criteria optimization on random forest for efficient human activity classification. *International Journal of Advanced Computer Science and Applications*, *10*(6).

Hu, Y.-C. (2007). Fuzzy integral-based perceptron for two-class pattern classification problems. *Information Sciences*, *177*(7), 1673–1686.

Hussein, A. Y., Falcarin, P., & Sadiq, A. T. (2021). Ids in iot using the iotid20 dataset with one-hot encoding and random forest [Open Access, peer-reviewed]. *Periodicals of Engineering and Natural Sciences*, *9*(3), 579–591. https://doi.org/10.21533/pen.v9i3.2204

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big data*, *1*(4), 215–226.

Kim, M.-G., & Desaire, H. (2024). Detecting the use of chatgpt in university newspapers by analyzing stylistic differences with machine learning. *Information*, *15*(6), 307.

Lei, L., Zhang, H., & Yang, S. X. (2023). Chatgpt in connected and autonomous vehicles: Benefits and challenges. *Intelligent Robots*, *3*, 144–147. https://doi.org/10.20517/ir.2023.08

Levy, J. J., & O'Malley, A. J. (2020). Don't dismiss logistic regression: The case for sensible extraction of interactions in the era of machine learning. *BMC medical research methodology*, *20*(1), 171.

Liao, L., Li, H., Shang, W., & Ma, L. (2022). An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *31*(3), 1–40.

Liu, X., Zhu, Y., Zhu, S., Liu, P., Liu, Y., & Yu, D. (2024). Evaluating moral beliefs across llms through a pluralistic framework. *arXiv preprint arXiv:2411.03665*.

Lodwich, A., Rangoni, Y., & Breuel, T. (2009). Evaluation of robustness and performance of early stopping rules with multi layer perceptrons. *2009 international joint conference on Neural Networks*, 1877–1884.

Lu, H.-m. (2020). Quasi-orthonormal encoding for machine learning applications. *CoRR*, *abs/2006.00038*. https://arxiv.org/abs/2006.00038

Norden, L., & Lerude, B. (2023). *States take the lead on regulating artificial intelligence* [Published: November 1, 2023, Last Updated: November 6, 2023, Accessed: 2024-09-09]. https://www.brennancenter.org/our-work/research-reports/states-take-lead-regulating-artificial-intelligence

Pinheiro, I., Moreira, G., Magalhães, S., Valente, A., Cunha, M., & dos Santos, F. N. (2024). Deep learning based approach for actinidia flower detection and gender assessment. *Scientific Reports*, *14*(1), 24452. https://doi.org/10.1038/s41598-024-73035-1

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Rezaeian Zadeh, M., Amin, S., Khalili, D., & Singh, V. P. (2010). Daily outflow prediction by multi layer perceptron with logistic sigmoid and tangent sigmoid activation functions. *Water resources management*, *24*, 2673–2688.

Sanchez-Medina, J. J. (2024). Sentiment analysis and random forest to classify llm versus human source applied to scientific texts. *arXiv preprint arXiv:2404.08673*.

Scikit-learn Developers. (2024). *Permutation importance* [Version 1.5]. Scikit-learn. https://scikit-learn.org/1.5/modules/permutation_importance.html#id2

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature

hashing (dissertation). https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426

Singh, A., Murzello, Y., Lee, H., Abdalla, S., & Samuel, S. (2024). Moral decision making: Explainable insights into the role of working memory in autonomous driving. *Machine Learning with Applications*, 100599.

Singh, A., Murzello, Y., Pokhrel, S., & Samuel, S. (2024). Classification of moral decision making in autonomous driving: Efficacy of boosting procedures. *Information*, *15*(9), 562.

Steed, R., & Williams, B. (2020). Heuristic-based weak learning for automated decision-making. *arXiv preprint arXiv:2005.02342*.

Takemoto, K. (2024). The moral machine experiment on large language models. *Royal Society Open Science*, *11*, 231393. https://doi.org/10.1098/rsos.231393

United Nations. (2024). Regional groups of member states [Accessed: 2024-11-25]. https://www.un.org/dgacm/en/content/regional-groups

Van Belle, V., & Lisboa, P. (2014). White box radial basis function classifiers with component selection for clinical prediction models. *Artificial intelligence in medicine*, *60*(1), 53–64.

Vida, K., Damken, F., & Lauscher, A. (2024). Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 1490–1501.

Vitolla, F., Raimo, N., Rubino, M., & Garegnani, G. M. (2021). Do cultural differences impact ethical issues? exploring the relationship between national culture and quality of code of ethics. *Journal of International Management*, *27*(1), 100823.

Wen, Z., Shi, J., Li, Q., He, B., & Chen, J. (2018). Thundersvm: A fast svm library on gpus and cpus. *Journal of Machine Learning Research*, *19*(21), 1–5.

Wiedeman, C., Wang, G., & Kruger, U. (2020). Modeling of moral decisions with deep learning. *Visual Computing for Industry, Biomedicine, and Art*, *3*(1), 27.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846. https://doi.org/https://doi.org/10.1016/j.patcog.2015.03.009

APPENDIX A

This research employed Python programming language (version 3.11.5) in Jupyter Notebook. Visual Studio Code and GoogleColab were used to write code in. Libraries, packages and modules used to conduct analyses present in this research include:

- Pandas (version 2.0.3)
- Numpy (version 1.26.4)
- Matplotlib.pyplot (version 3.7.2 of matplotlib)
- ThunderSVM (0.1)
- Seaborn (version 0.12.2)
- Several packages from `sci-kit learn`, including `train_test_split`, `accuracy_score`, `confusion_matrix`, `classification_report`, `LogisticRegression`, `RandomForestClassifier`, `KFold`, `StratifiedKFold`, `permutation_importance`, `RandomizedSearchCV` & `Permutation_importance`.

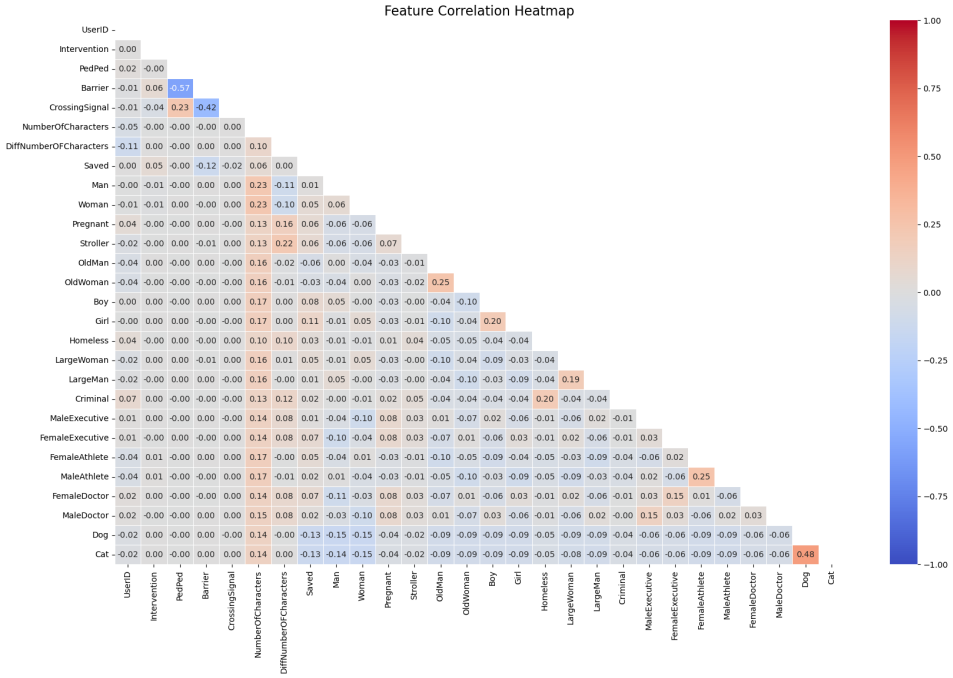


Figure 19: Feature correlation heatmap df_total dataset

APPENDIX B

The feature correlation heatmaps per dataset.

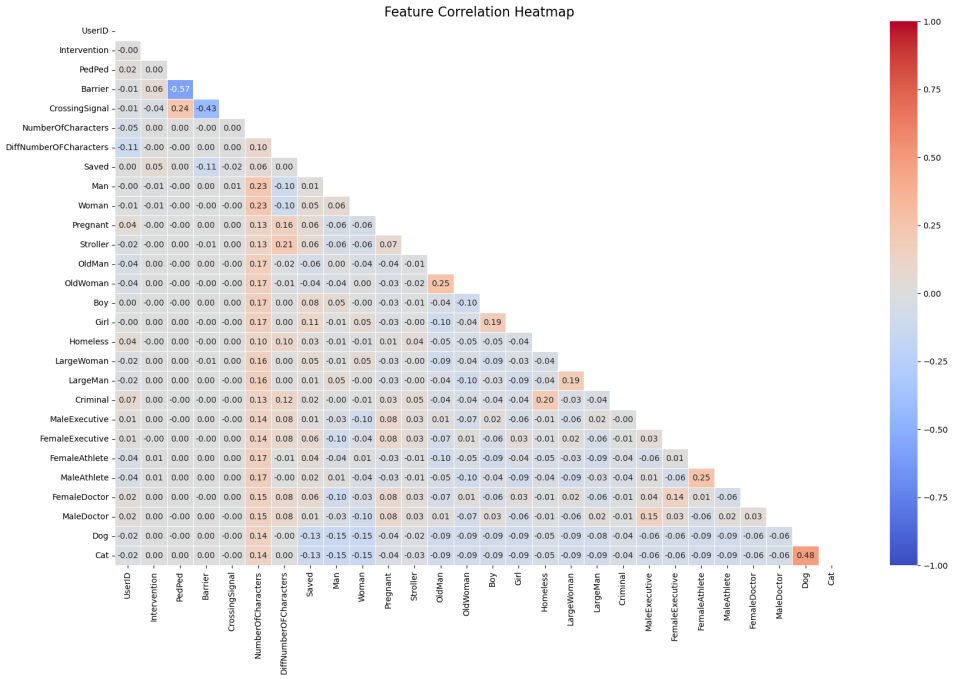


Figure 20: Feature correlation heatmap df_non-western dataset

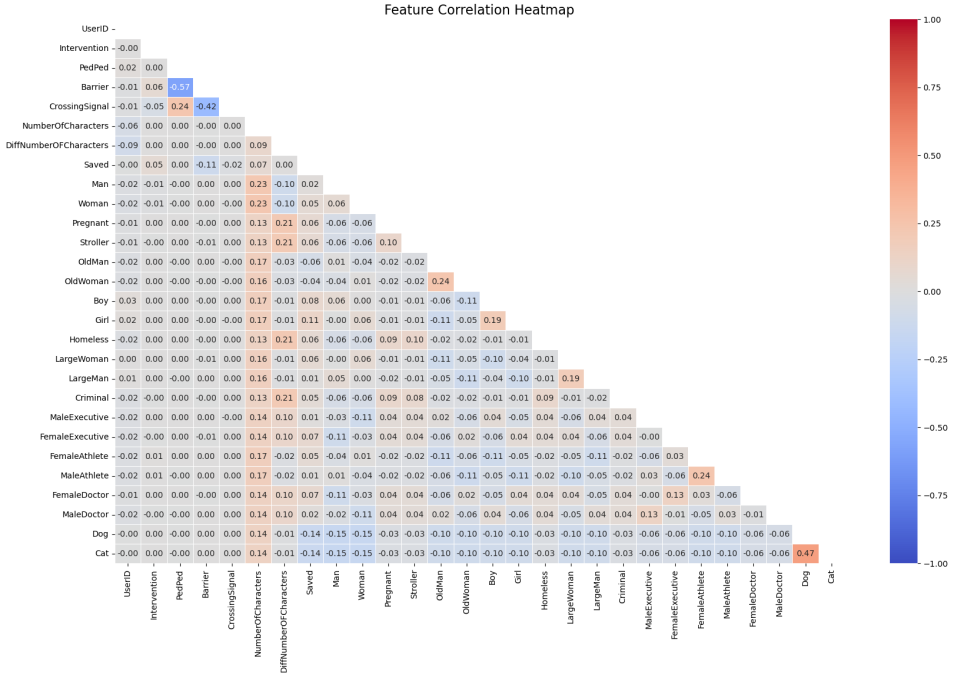


Figure 21: Feature correlation heatmap df_deleted dataset

APPENDIX C

Table 6 shows each of the features that describe the scenarios. The feature explanation has been obtained from Awad et al. (2018) OSF ReadMe page.

Feature Name	Explanation
`Intervention`	Represents the decision of the AV (STAY or SWERVE) that would lead to this outcome [0: the character would die if the AV stays, 1: the character would die if AV swerves]. This is not the actual decision taken by the user, but rather a part of the structural characterisation of the scenario.
`PedPed`	Every scenario has either pedestrians vs. pedestrians or pedestrians vs. passengers (or passengers vs. pedestrians). This column provides information about not just this outcome, but about the combination of both outcomes in the scenario; whether the scenario pits pedestrians against each other or not [1: pedestrians vs. pedestrians, 0: pedestrians vs. passengers (or vice versa)].
`Barrier`	Structural column which describes whether the potential casualties in this outcome are passengers or pedestrians [1: passengers, 0: pedestrians]. This column was used to calculate PedPed.
`CrossingSignal`	Structural column which represents whether there is a traffic light in this outcome, and light colour if yes [0: no legality involved, 1: green or legally crossing, 2: red or illegally crossing]. Every scenario that has pedestrians vs. pedestrians (i.e. PedPed=1) features one of three legality-relevant characterisations: a) the pedestrians on both sides are crossing with no legal complications, b) one group is crossing legally (on a green light), while the other is crossing illegally (on a red light), and c) vice versa. Every scenario that has pedestrians vs. passengers (i.e. PedPed=0) features also one of three legality-relevant characterisations: a) the pedestrians are crossing with no legal complications, b) the pedestrians are crossing legally (on a green light), and c) pedestrians are crossing illegally (on a red light). There are no legality concerns for passengers.
`Number OfCharacters`	Takes a value between 1 and 5, the total number of characters in this outcome. This is the sum of numbers in the 20 character columns. It also represents the number of characters who will be saved or killed based on “Saved" value.

Feature Name	Explanation
DiffNumber OFCharacters	Takes a value between 0 and 4; difference in number of characters between this outcome and the other outcome.
Man	Indicates the number of male adult characters in the outcome.
Woman	Indicates the number of female adult characters in the outcome.
Pregnant	Indicates the number of a pregnant woman in the outcome.
Stroller	Indicates the number of a stroller in the outcome.
OldMan	Indicates the number of elderly male characters in the outcome.
OldWoman	Indicates the number of elderly female characters in the outcome.
Boy	Indicates the number of male child characters in the outcome.
Girl	Indicates the number of female child characters in the outcome.
Homeless	Indicates the number of a homeless characters in the outcome.
LargeWoman	Indicates the number of a larger-bodied female characters in the outcome.
LargeMan	Indicates the number of a larger-bodied male characters in the outcome.
Criminal	Indicates the number of a criminal characters in the outcome.
MaleExecutive	Indicates the number of a male executives in the outcome.
FemaleExecutive	Indicates the number of a female executives in the outcome.
FemaleAthlete	Indicates the number of a female athletes in the outcome.
MaleAthlete	Indicates the number of a male athletes in the outcome.
FemaleDoctor	Indicates the number of a female doctors in the outcome.
MaleDoctor	Indicates the number of a male doctors in the outcome.
Dog	Indicates the number of a dogs in the outcome.
Cat	Indicates the number of a cats in the outcome.

Feature Name	Explanation
`AttributeLevel`	Is dependent on the scenario type. Each scenario type has two levels: +Gender: [Males: characters are males, Females: characters are females] +Age: [Young: characters in this outcome are younger (Boy/Girl + Man/Woman) than in the other outcome, Old: characters in this outcome are older (Elderly Man/Woman and Man/Woman)]. +Fitness: [Fit: characters in this outcome are more fit (Male/Female Athlete and Man/Woman), Fat: characters in this outcome are less fit (Large Man/Woman and Man/Woman)]. +Social Value: this was changed in the analysis to "social status" instead, and the characters Male/Female Doctor and Criminal were filtered out [High: characters in this outcome have higher social status (Male/Female Executives and Man/Woman), Low: characters have a lower social status (Homeless and Man/Woman)]. +Species: [Hoomans: characters in this outcome are humans (all but Dog/Cat), Pets: characters in this side are pets (Dog/Cat)]. +Utilitarian: [More: there are more characters in this outcome, Less: there are fewer people in this outcome]. In fact, the characters on the "More" side are the same characters on the "Less" side, in addition to at least one more character.
ScenarioTypeStrict	These two columns have 6 values, corresponding to 6 types of scenarios. These are: "Utilitarian","Gender", "Fitness", "Age", "Social Value", and "Species".

APPENDIX D

The MLP model utilized in this researh has two dense (fully connected) hidden layers, each containing 64 neurons. Both layers use the ReLU activation function. Batch normalization is applied after each hidden layer to normalize the input to each layer. The output layer is a single neuron with a sigmoid activation function. Sigmoid is well-suited for binary classification, as it outputs a probability value between 0 and 1 (Hu, 2007). The model is compiled using the Adam optimizer with an exponentially decaying learning rate starting at 0.0005. Binary cross-entropy is employed as the loss function. The model deviates from that of Wiedeman et al. (2020) in its choice evaluation metric, as here Recall is added as an extra metric, on top of accuracy. An early stopping criterion is added, as early stopping helps prevent overfitting and can reduce the runtime of the models drastically (Lodwich et al., 2009).