PREDICTING HUMAN DECISIONS IN AUTONOMOUS VEHICLE CRASHES: THE ROLE OF THE AVAILABILITY BIAS

K-PROTOTYPE CLUSTERING, BINARY LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST AND NEURAL NETWORKS

TIMO KLEIN

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

2040880

COMMITTEE

dr. Klincewicz
Mr. Mohebbi

LOCATION

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

DATE

December 2nd, 2024

WORD COUNT

8778

ACKNOWLEDGMENTS

I would like to sincerely thank my supervisor, dr. Klincewicz, for his guidance, expertise, and encouragement throughout my research. Your feedback and support have been invaluable in shaping the direction of my thesis. I am forever grateful to my family for their endless love and encouragement. To my parents, older sister and twin brother, your belief in me has been a constant source of motivation during my entire academic career. I would also like to thank my friends for being a source of comfort and joy throughout this process. Without the fun, laughter, and the wonderful moments with you in the library, I would never have been able to accomplish this work. Ironically, during the writing of this thesis on road traffic accidents, I experienced one myself. I was hit by a car but, fortunately, only broke my collarbone. Even during this time, my supervisor was deeply involved, and I received incredible support from my family and friends. For this, I am immensely grateful. Thank you to everyone who has contributed to this journey.

PREDICTING HUMAN DECISIONS IN AUTONOMOUS VEHICLE CRASHES: THE ROLE OF THE AVAILABILITY BIAS

K-PROTOTYPE CLUSTERING, BINARY LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST AND NEURAL NETWORKS

TIMO KLEIN

Abstract

This study tries to predict human moral decisions regarding whether to save or kill individuals in fatal road traffic accidents involving autonomous vehicles (AVs). Using data from the Moral Machine platform, the research examines the effects of prior knowledge about AVs and experiences with road traffic accidents on these moral decisions. A range of models, including binary logistic regression, decision trees, random forests, and neural networks, were employed and compared for their predictive performance. Analysis using permutation feature importance and partial dependence plots revealed the underlying relationships between important features. Among the models, the neural network demonstrated the highest accuracy at 71.4%. However, no evidence was found to support the hypothesis that prior knowledge about AVs and experiences with road traffic accidents directly impacts moral decisions. The study did reveal that people with higher skill levels were less likely to endorse utilitarian AVs compared to those with lower skill levels and that people exposed to road traffic accidents involving passengers were less likely to favor utilitarian AVs. The findings of this paper add to the understanding of human decision-making and highlight the potential impact of cognitive biases on the design and functionality of AV algorithms.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The data used in this study is data from the Moral Machine website (Awad et al., 2024). The data from the Moral Machine platform is publicly available and acquired through the Open Science Framework: Moral Machine (Awad, 2021). Given its public availability, I am allowed to use this data in my research project, provided I adhere to any applicable terms and conditions set by the data provider. The Moral Machine website gathers information on people's judgments regarding decision-making in moral dilemmas and does thus involve collecting data from human participants. The data is owned by the Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. (MPG). They are responsible for processing and storing the data in compliance with the EU General Data Protection Regulation (GDPR). They process personal data only after obtaining the user's consent, as outlined in their GDPR compliance statement. The original owner of the data used in this thesis retains ownership of the data and code during and after the completion of this thesis. All the figures belong to the author, except for Figure 3 and Figure 6. These consist of a screenshot from the Moral Machine website and a figure from the paper by Wiedeman et al. (2020). They are appropriately cited according to APA citation style. The thesis code can be accessed through the GitHub repository following the link: GitHub. This code can be used to replicate the findings of this thesis. No code is used from another study or individual. The programming software Visual Studio Code and Google Colab and the programming language Python (v3.12.7) were used to write the code (Microsoft, 2024; Google, 2024; Foundation, 2024). Several libraries were utilized, and their names along with the corresponding version and sources can be found in Appendix A. ChatGPT, developed by OpenAI, was used to assist in writing and debugging code, as well as to improve the clarity and coherence of the text in the thesis (OpenAI, 2024).

2 INTRODUCTION

2.1 *Background*

Artificial intelligent (AI) systems are increasingly being used to make decisions in critical areas such as hiring, criminal justice and healthcare (Dattner et al., 2019; Sushina and Sobenin, 2020; Shaheen, 2021). This rapid development of AI has led recent research to study the human moral decisions that should be implemented in the behavior of AI algorithms (Frank et al., 2019). However, building morally just algorithms based on people's moral decisions raises the issue of biases and how algorithms can

adopt or even amplify the biases present in human moral decision-making (Frank et al., 2019; Barocas and Selbst, 2016; Rich and Gureckis, 2019).

Developing fair AI algorithms without biases requires a deeper understanding of how people want AI systems to make decisions in critical areas (Awad et al., 2024). The Moral Machine website seeks to provide this insight in the context of autonomous vehicles (AVs) by collecting data on human decision-making in scenarios involving self-driving cars (Awad et al., 2024). The experiment asked respondents to make decisions in various ethical dilemmas where a self-driving car had to choose between saving and killing different groups of characters in fatal car accident scenarios. The present paper uses the resulting dataset to examine the presence of the availability bias in human decision-making, where the availability bias can be described as a decision-making heuristic where individuals rely on relevant, recent information that come to mind easily to make quick decisions and judgments (Tversky and Kahneman, 1973). The study demonstrates the presence of the availability bias by looking how people's decisions are biased due to the level of knowledge people have about AVs and due to past road traffic accidents.

2.2 *Problem Statement*

The goal of this research is to predict human moral decisions regarding whether to save or kill individuals in fatal road traffic accidents involving AVs.

The focus of this paper is on the effect of prior knowledge about AVs and prior road traffic accidents on human moral decisions in road traffic accidents involving AVs. To quantify prior knowledge, the Frontier Technology Readiness Index (FTRI) will be used (UNCTAD, 2023). This index assesses a country's readiness to adopt, adapt, and implement frontier technologies, such as AVs (UNCTAD, 2023). To represent fatal road traffic accidents, data from the World Health Organization will be used, which includes estimates of road traffic deaths among females, males, passengers, pedestrians, and the total population in a country (WHO, 2020; WHO, 2021). The statistics on the road traffic deaths, as well as the statistics on the frontier technology readiness index, will in this paper often be referred to as prior information about AVs. All features serve as proxy for the availability bias and are used to predict human moral decisions regarding whether to save or kill individuals in fatal road traffic accidents involving AVs.

Additionally, this paper looks if the effect of the number of lives that can be saved in a scenario on the decision of people in that scenario is dependent on prior knowledge about AVs and past road traffic accidents.

In the remainder of this paper, AVs programmed to save the largest number of road users will be referred to as utilitarian AVs.

2.3 *Scientific Relevance*

Limited research has been conducted on the effect of the availability bias on moral decision-making in AVs, and the only two existing studies provide contradictory results (Zhu et al., 2022; Othman, 2023). While both papers found that prior information about AVs significantly impacts the public's moral judgments, they disagree on how prior knowledge influences the preference for utilitarian AVs. The present study aims to fill this gap in the literature by investigating this effect where prior studies reported contradictory results. Moreover, no study yet explored the availability bias in the context of the Moral Machine platform (Awad et al., 2024). The present study utilizes the large-scale Moral Machine dataset leading to a significantly larger subject pool compared to previous studies. While research on biases in AVs has grown substantially in recent years, this study thus examines the under-researched availability bias and utilizes a significantly larger dataset. It is moreover unique in its methodology, employing both unsupervised and supervised machine learning techniques to explore patterns indicative of human biases.

2.4 *Societal Relevance*

As said before, algorithms can adopt cognitive biases present in human moral decision-making (Frank et al., 2019; Barocas and Selbst, 2016; Rich and Gureckis, 2019). Research has shown that if systems carry biases, they may perpetuate or even amplify existing societal inequalities (Barocas & Selbst, 2016). This can produce unequal outcomes that disadvantage individuals based on factors such as race, gender, or socioeconomic status. (Žliobaitė, 2017). Moreover, decision-making biases may introduce distortions in the decision making process of individual's (Chira et al., 2008). These distorted, irrational biases can make it hard for machines to learn a consistent ethical framework for the behavior of their algorithms and to apply this consistent framework when faced with ethical decisions, such as saving or sacrificing road users. (Frank et al., 2019). Ensuring AI algorithms are fair and consistent is crucial to protect individuals from discrimination and to promote equity (Binns, 2018).

This paper aims to improve the understanding of how humans make choices and how biases in these decisions may influence the design and functioning of AI algorithms in AVs. By exploring the under-researched area of the availability bias, we can contribute to existing knowledge

and help prevent future algorithms from being contaminated by negative human decision-making biases.

2.5 *Research Questions*

RQ *What is the influence of the availability bias on human moral decisions whether to save or kill characters in fatal road traffic accidents involving autonomic vehicles?*

Sub-questions:

SQ1. *How do the performance of the machine- and deep-learning models binary logistic regression, decision tree, random forest, and neural networks compare when predicting the moral decision whether to kill or save characters in the Moral Machine scenarios?*

This question focuses on obtaining the highest model performance in this study. A high model performance means the availability bias can accurately predict moral decisions in autonomic vehicles.

SQ2. *Are the estimated number of road traffic deaths for males, females, pedestrians, and passengers, as well as the statistics on the frontier technology readiness index, predictive of the moral decision of whether to save or kill characters in the Moral Machine scenarios?*

Using permutation feature importance, the contribution of each feature to the model's performance will be measured.

SQ3. *Is the preference for utilitarian AVs affected by the availability bias?*

The feature importance of the interaction term between the number of lives saved and the number of road traffic deaths for males, females, pedestrians, and passengers, as well as the statistics on the frontier technology readiness index, will be measured. Additionally, a partial dependence plot will visualize how the relationship between the number of lives saved and the likelihood of saving characters varies depending on prior information about AVs.

3 LITERATURE REVIEW

In this section, we will discuss the most relevant literature related to the research in this paper. First, we will give a quick overview of the most influential work on biases focusing on the availability bias. Second, we will discuss the growing body of scientific studies investigating human decision-making in AV accidents, including the literature on the Moral Machine. Third, we will look at papers that directly apply the availability bias to decision-making in AVs. Lastly, based on the literature review, we will formulate our hypotheses for the research questions.

3.1 *The Availability Bias*

The influence of cognitive biases on human moral decisions has been researched for a long time, starting with Tversky and Kahneman (1973). Since then, numerous studies have refined our understanding of biases such as the prospect theory, the hot hand fallacy, and hindsight bias (Kai-Ineman and Tversky, 1979; Gilovich et al., 1985; Kahneman, 2003; Bazerman and Moore, 2012).

In 1973 and 1974, Tversky and Kahneman introduced key heuristics used in judgments under uncertainty, including the availability bias (Tversky and Kahneman, 1973; Tversky and Kahneman, 1974). This bias can be described as a decision-making heuristic where individuals rely on relevant, recent information that come to mind easily to make quick decisions and judgments (Tversky and Kahneman, 1973). Lichtenstein et al. (1978) and Combs and Slovic (1979) provide insights into how the availability bias affects people's perceptions of lethal events. Lichtenstein et al. (1978) found that participants overestimated death frequencies from memorable events like accidents and homicides. They link this to the availability bias, where the ease of recall influences frequency estimation. Combs and Slovic (1979) showed that US newspapers over-reported catastrophic deaths, also reinforcing biases in human risk perception. Additionally, Schwarz et al. (1991) highlighted that the ease of retrieval plays a significant role in probability assessments. Although these studies were conducted long before AVs gained widespread attention, their findings suggest that public perception and decision-making in AV accidents could be influenced by the availability bias due to media coverage and ease of recall.

3.2 *Decision-Making in AV Accidents: The Moral Machine*

3.2.1 *Decision-Making in AV Accidents*

Recent literature increasingly explores decision-making in AV accidents due to the potential benefits of AVs, such as improved transportation efficiency, safety, and reduced accidents caused by humans (Wang et al., 2020). A central theme in current research is understanding how people believe AVs should act in accident scenarios. Studies collected responses from participants worldwide, including Sweden, Japan, Germany, and the United Kingdom (Habla et al., 2024; Bodenschatz et al., 2021; Takaguchi et al., 2022). Methods ranged from surveys with numerous moral dilemmas, conducted by Habla et al. (2024), to immersive virtual reality used by Faulhaber et al. (2019) and Sütfeld et al. (2017). Bergmann et al. (2018) found participants were willing to sacrifice their lives to save others, while Habla et al. (2024) noted that compliance with social norms and age influenced decisions. Takaguchi et al. (2022) found UK respondents favored saving more lives or family members, while Japanese respondents preferred saving pedestrians. Many studies also explored preferences for utilitarian or self-protective algorithms (Bonnefon et al., 2016; Wang et al. (2020); Ng, 2024; Liu and Liu, 2021; Bodenschatz, 2024). Though people acknowledged that sacrificing themselves to save a higher number of road users was most ethical, they preferred themselves to ride in self-protective AVs (Wang et al., 2020; Bonnefon et al., 2016; Liu and Liu, 2021; Ng, 2024).

3.2.2 *The Moral Machine*

The Moral Machine experiment and papers discussing the results of this experiment have made significant contributions to the literature on decision-making in AV accidents (Awad et al., 2018; Awad, Dsouza, Shariff, et al., 2020; Awad, Dsouza, Bonnefon, et al., 2020; Noothigattu et al., 2018; Shariff et al., 2017; Bonnefon et al., 2016). The authors of Awad et al. (2018) developed the Moral Machine platform, which gathered responses from millions of participants across 233 countries on ethical decisions involving AV accidents. Awad et al. (2018) and Awad, Dsouza, Shariff, et al. (2020) examined variations in moral preferences for AV algorithms based on cultural, demographic, and socioeconomic factors, highlighting cultural and individual differences in ethical frameworks for AV decision-making systems. The authors of Awad et al. (2018) wrote several other papers discussing the Moral Machine platform and the challenges and implications of AVs (Shariff et al., 2017; Awad, Dsouza, Bonnefon, et al., 2020; Bonnefon et al., 2016).

Few studies have used the Moral Machine dataset with machine learning or deep learning models. Noothigattu et al. (2018) applied machine learning techniques to develop models that learn participant preferences, which were combined into an overarching model to infer societal preferences for ethical dilemmas. Wiedeman et al. (2020) also built models to learn participant preferences in the Moral Machine, comparing a hierarchical Bayesian model by Kim et al. (2018), with a machine learning model and a deep neural network. Their machine learning model achieved a maximum accuracy of 0.66, while the deep neural network obtained accuracies ranging from 0.60 to 0.75, suggesting that the deep neural network can effectively model human ethics. The hierarchical Bayesian model by Kim et al. (2018) also demonstrated accuracies between 0.65 and 0.75, but the deep learning model by Wiedeman et al. (2020) outperformed it when the data distribution was skewed. Agrawal et al. (2019) compared a multilayer feedforward neural network with more interpretable models and found that the neural network outperformed simpler models for datasets with over 100,000 data points. The highest accuracy achieved by their neural network was equal to 0.774.

3.3 *Availability Bias in AV Accidents*

A growing body of studies investigates biases in decision-making in AVs, focusing on action bias, status quo bias, agency bias, conformity bias, self-protective bias, and algorithmic bias (Frank et al., 2019; Lim and Taeihagh, 2019; Mayer et al., 2021; Sui, 2023).

To the best of my knowledge, only two papers studied the availability bias in decision-making in AV accidents, those by Zhu et al. (2022) and Othman (2023). Only one paper studied biases in the context of the moral machine platform, namely Frank et al. (2019). No study has yet explored the availability bias in the context of the moral machine platform. Both Zhu et al. (2022) and Othman (2023) examine how people's moral judgments are affected by prior information about AVs. Zhu et al. (2022) focus on how expertise influences preferences for ethical algorithms in AVs, while Othman (2023) examines the impact of prior knowledge about AVs and AV accidents on public perception. Both papers found that prior information about AVs, both due to expertise and AV accidents, significantly impacts public moral judgments (Zhu et al., 2022; Othman, 2023). Both papers also found that preferences for utilitarian or self-protective AVs, as discussed in subsection 3.2.1, are affected by prior knowledge about AVs. Zhu et al. (2022) suggest that people in the AV industry are less likely to endorse utilitarian algorithms compared to outsiders. People in the AV industry, and thus more expertise, prioritized

AVs that protect the passenger. Othman (2023) partly contradicts Zhu et al. (2022), showing that respondents with prior knowledge about AVs are more likely to endorse utilitarian algorithms. However, they also showed that exposure to traffic accidents increased people's concern about personal safety, leading to decreased support for utilitarian algorithms.

3.4 *Hypotheses based on the literature*

SQ1 - Based on the existing literature, it can be expected that the deep neural network will demonstrate a higher predictive accuracy in forecasting moral decisions regarding AV accidents compared to simpler models like binary logistic regression and a decision tree. This is supported by the findings from Wiedeman et al. (2020) and Agrawal et al. (2019), who demonstrated that deep learning models outperform simpler models in understanding moral decisions in the Moral Machine dataset. **SQ2 -** It is also reasonable to expect that the estimated number of road traffic deaths for different groups and the frontier technology readiness index will significantly contribute to the model's predictions. Both Zhu et al. (2022) and Othman (2023) namely found that prior information about AVs, both due to expertise and AV accidents, significantly influenced people's moral judgments. **SQ3 -** The availability bias will likely affect preferences for utilitarian autonomous vehicles (AVs), with people more often exposed to AV accidents more likely to save less number of lives (Zhu et al., 2022). However, it is unclear how prior knowledge of AVs will affect the relationship between the number of lives saved and the likelihood of saving characters, as the literature on this topic is inconclusive.

RQ - The overall hypothesis is that the availability bias will significantly influence human moral decisions regarding whether to save or kill characters in fatal road traffic accidents involving AVs, particularly within the context of the Moral Machine scenarios. This hypothesis is supported by established literature suggesting that decision-making can be significantly influenced by the availability bias due to media coverage and the ease with which people can recall information (Lichtenstein et al., 1978; Combs and Slovic, 1979; Schwarz et al., 1991). Recent studies further support this hypothesis, demonstrating that prior information about AVs has a significant impact on the public's moral judgments (Zhu et al., 2022; Othman, 2023).

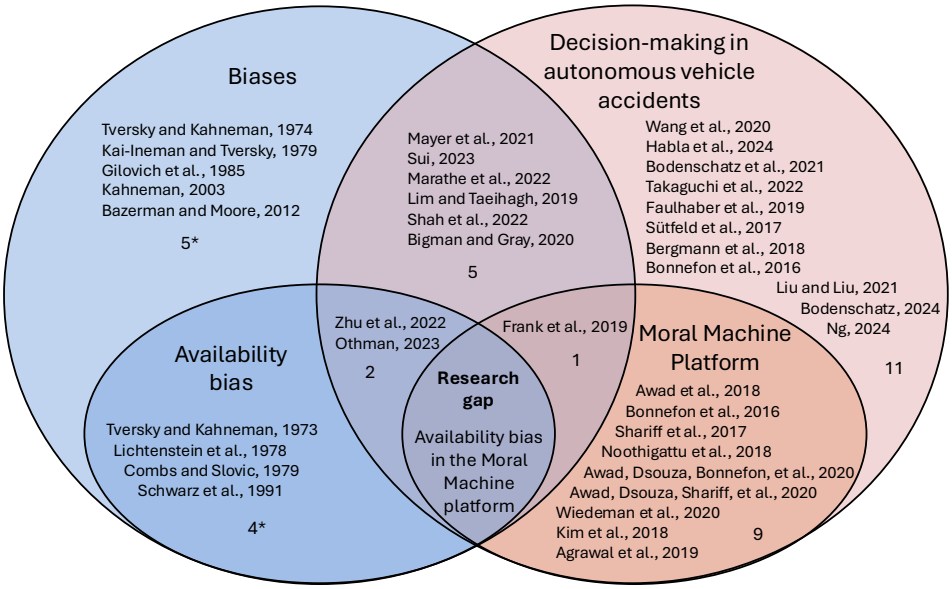


Figure 1: Literature Review Structure (*most important papers).

4 METHOD

This section outlines the methodology with the flowchart in Figure 2. It discusses de used dataset, the preprocessing steps such as data cleaning and feature engineering, and the models used in this study.

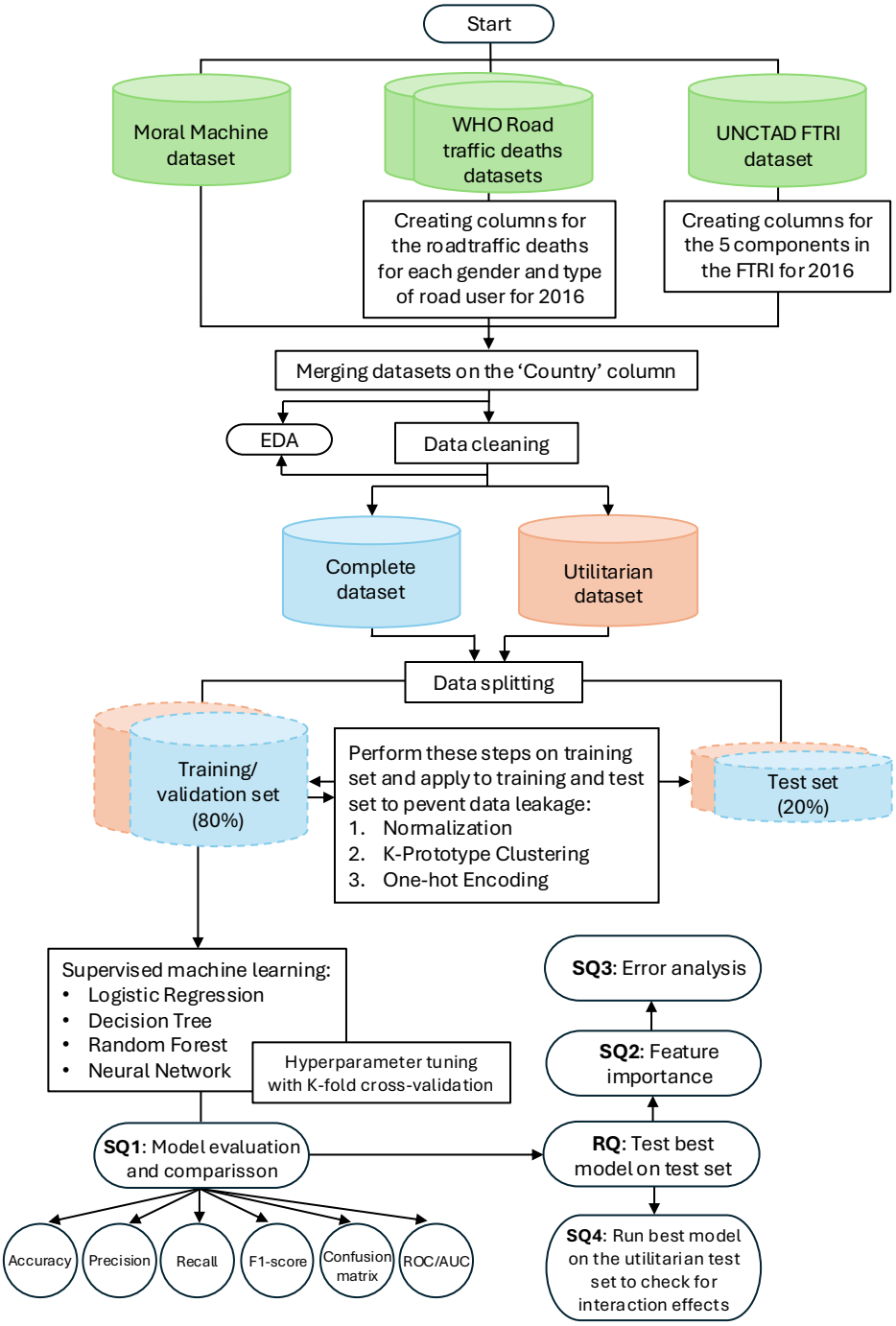


Figure 2: Flowchart of the research methodology described in section 4.

4.1 *Data description*

The data used in this study is data from the Moral Machine (Awad, 2021), an online platform created by researchers from MIT to gather data on human moral decisions regarding AVs (Awad et al., 2024). This dataset will be referred to as the 'Moral Machine' dataset. The experiment asks respondents to make decisions in ethical dilemmas where a self-driving car must choose between saving and killing different groups of individuals in fatal car accident scenarios. Each session on the Moral Machine platform consists of 13 scenarios, each with a choice between two outcomes. The car can either stay in its current lane, killing the group of individuals there, or swerve to another lane, killing a different group. Figure 3 shows one of these scenarios. When a respondent makes a choice, it is recorded in a database, with one row for the chosen outcome and another for the alternative outcome. The original dataset used in this paper consists of over 70 million rows. Some columns include features representing respondents, such as a unique user ID and session ID. Other features describe characteristics of the outcome, such as the number of characters, whether pedestrians are crossing a traffic light, and the count of each character type. The target variable 'Saved' resembles the actual decision made by the user. It is 1 if the user decided to save the characters in an outcome and 0 if they decided to kill them.

To enrich the Moral Machine dataset, three additional datasets from the World Health Organization are used (World Health Organization, 2024). The 'Road Traffic Deaths Gender' dataset contains data from 183 countries on the number of road traffic deaths, in absolute numbers and per 100,000 people, disaggregated by gender (WHO, 2021). The 'Road Traffic Deaths User' dataset includes data from 117 countries on the distribution of road traffic deaths by type of road user (WHO, 2020). The third dataset, the 'FTRI' dataset, presents the frontier technology readiness index (FTRI) for 166 countries, developed by UNCTAD (UNCTAD, 2023). The statistics include 'ICT', 'Skills', 'Research and Development', 'Industry activity', 'Access to finance', and an 'Overall index'.

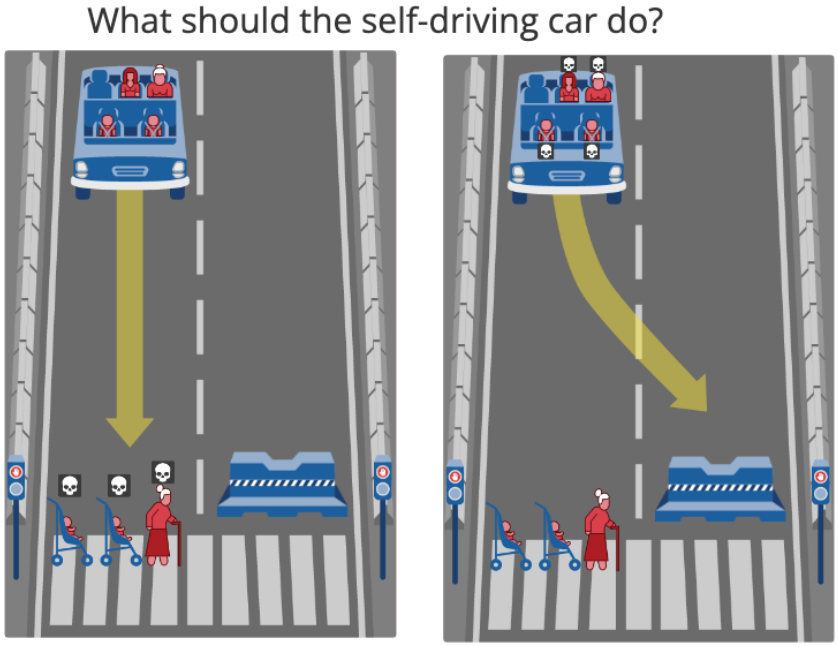


Figure 3: Example of a scenario from the Moral Machine platform (Awad et al., 2024). The respondents must choose between saving or killing the characters on the left or the right.

4.2 *Preprocessing*

4.2.1 *Data Cleaning*

From the 'Road Traffic Deaths Gender' dataset and the 'Road traffic Deaths User' dataset, features with the estimated road traffic death rate (per 100,000 population) for males, females, passengers, pedestrians and total population per country are selected or engineered. From the 'TFRI' dataset, all aforementioned statistics on the frontier technology readiness index are kept. For each feature and country, only data from 2015 and 2016 are kept as a proxy for the availability bias. When people unconsciously rely on the availability bias, they namely rely on recent information that comes to mind easily (Tversky and Kahneman, 1973). Since the Moral Machine platform went live in 2016, the average death rate from 2015 and 2016 serves as a suitable proxy. The variables in the 'Country' column are converted to ISO codes for easy merging with the Moral Machine dataset later. All missing values in the datasets are deleted.

The Moral Machine dataset is processed in chunks to ensure efficient memory usage. Only complete sessions with 26 outcomes, and thus 13 scenarios, are kept. By keeping only complete sessions, the effect of previous choices in the session is considered. According to Abrahamyan et al. (2016),

prior choices can namely play an important role when making decisions under uncertainty. Abrahamyan et al. (2016) found that human decisions, even in straightforward two-option decision tasks like those in the Moral Machine, are influenced by prior decisions. Each chunk is merged with the three aforementioned datasets ('Road Traffic Deaths Gender,' 'Road Traffic Deaths User,' and 'TFRI') based on the 'Country' column. The columns from these datasets will serve as features representing the availability bias. Unnecessary columns are deleted, and all missing values, duplicate rows, and rows with a unique 'ResponseID' are removed. Since 'ResponseID' represents a scenario identifier, it should always correspond to another row associated with the same scenario. After these data cleaning steps, the cleaned chunks are combined to create the final dataset for modeling, referred to as the "complete" dataset. This "complete" dataset contains 1,405,234 rows and 39 features. A detailed description of the final features is provided in Appendix B.

An additional dataset is created, including only "utilitarian" scenarios, hereinafter referred to as the "utilitarian" dataset. While most scenarios in the "complete" dataset have an equal number of characters in both outcomes, the "utilitarian" dataset contains only scenarios with a different number of character in both outcomes. In these scenarios, the characters are the same in both outcomes, with at least one additional character on one side. The "utilitarian" dataset contains the same number of rows as the "complete" dataset.

Exploratory data analysis (EDA) was conducted both before and after data cleaning to identify potential missing values, outliers, distributions, frequency counts, correlations, and interactions. All tables and graphs from the EDA are shown in Appendix G.

4.2.2 *Feature Engineering (K-Prototype Clustering)*

As described in subsection 4.2.1, three datasets from the World Health Organization are used to engineer features serving as proxies for the availability bias (World Health Organization, 2024). These features include the estimated number of road traffic deaths (per 100,000 population) for females, males, passengers, pedestrians, and the total population, as well as statistics on the frontier technology readiness index. For the "utilitarian" dataset, twelve additional features are engineered, including the number of lives saved and 11 interaction features between the proxies for availability bias and the number of lives saved. The number of lives saved ranges from minus four (saving one character instead of five) to four (saving five characters instead of one).

Normalization, one-hot encoding, and K-prototypes clustering were applied after splitting the data into training and test sets. The procedures

were carried out on the training set and then applied to both the training and test sets to prevent data leakage, which could lead to overestimated model accuracy (C. Yang et al., 2022). Details regarding the data split can be found in subsection 4.3.2. The numerical columns in both the "complete" and "utilitarian" dataset are normalized before K-Prototypes Clustering, which increases accuracy for classification and clustering (Alam, 2020). After clustering, the categorical columns are one-hot encoded, which often leads to higher accuracy compared to label encoding for various models (Gong and Chen, 2022).

Ashraf et al. (2021) uses an unsupervised machine learning model to identify patterns in contributing factors of AV-involved crashes. Since the Moral Machine scenarios also involve multiple factors describing AV-related incidents, this study also applies an unsupervised approach, specifically the K-Prototype Clustering method introduced by Huang (1997). K-Prototype Clustering extends k-means to handle mixed numeric and categorical data by assigning data points to clusters based on proximity to prototypes and updating the prototypes accordingly (Huang, 1997). This method allows clustering of both numerical and categorical features. The cluster labels obtained are added as features for the supervised models. To determine the optimal number of clusters, the elbow method was applied and visualized through an elbow curve. For each value of k, a K-Prototypes model was initialized using the Huang method (Huang, 1997). Due to computational constraints, the elbow method was applied to a subset of the training set. Figure 4 and Figure 5 show the relationship between the number of clusters and the associated cost for both datasets. For the "complete" dataset, a distinct "elbow" point is visible at three clusters, which was selected as the optimal number. For the "utilitarian" dataset, the decrease in cost slows down after six clusters. The K-Prototype Clustering algorithm, with three and six clusters, was then fit on the entire training sets and used to predict clusters for both sets. The predicted cluster labels were added as new columns to both training and test sets.

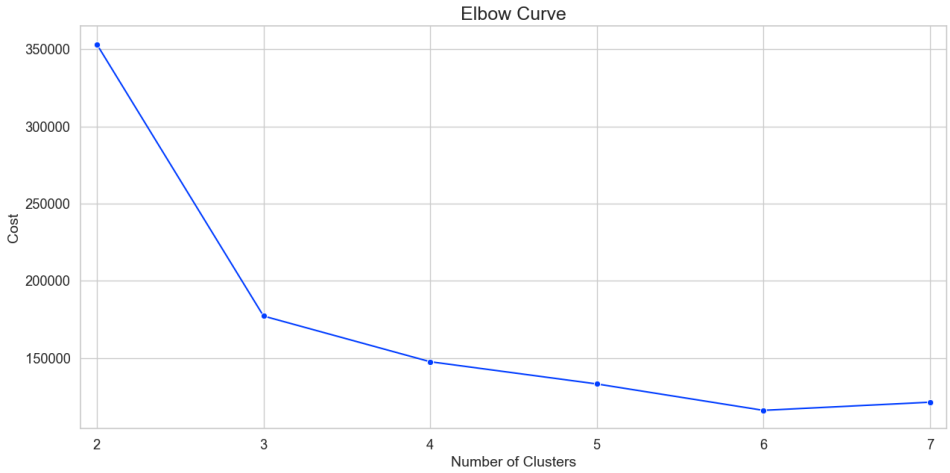


Figure 4: Elbow Curve for the "complete" dataset

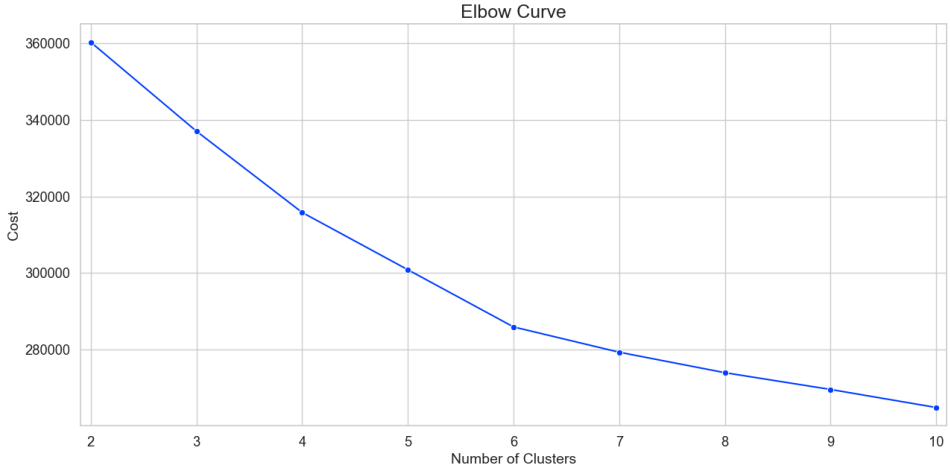


Figure 5: Elbow Curve for the "utilitarian" dataset

4.3 *Modeling*

To predict the binary target variable whether people will save or kill characters in the Moral Machine scenarios, various supervised machine- and deep-learning algorithms will be used. These algorithms include a binary logistic regression, decision tree, random forest and neural network.

4.3.1 *Hyper-parameter Tuning*

A random search is utilized to tune the hyper-parameters and optimize the performance of each model. Random search has been shown to perform

as effectively as grid search while achieving results more efficiently by sampling fewer sets of parameters from a broader search space (Bergstra and Bengio, 2012). The size of the search space varies per model, but from these search spaces, 25 sets of parameters are randomly tested on each model.

4.3.2 *Data Splitting and K-fold cross-validation*

The data is randomly split in a training/validation set and a test set. The training/validation set consists of 80% of the data and the test set of 20%, since this is a commonly used ratio (Joseph, 2022). The performance of each of the 25 candidate set of parameters is evaluated using K-fold cross-validation on the training/validation set. K-fold cross-validation is very popular for evaluating the performance of classification algorithms as well as comparing the performance between two or more classification algorithms (Wong, 2015). Given that this study focuses on comparing the performance of various models, K-fold cross-validation is thus well-suited for this purpose. The method evaluates a classification algorithm by dividing a dataset into k disjoint folds of approximately equal size (Wong, 2015). The training/validation set in this paper is divided into three disjoint folds for each learning algorithm. Each algorithm trains on two (k-1) folds and tests on the remaining fold, repeating this process for each fold. The performance is then assessed by averaging the evaluation metrics across all k iterations (Wong, 2015).

4.3.3 *Performance Evaluation*

The evaluation metrics used in this paper include accuracy, precision, recall and F1-score. These metrics are used to judge and compare the performance of the competing models, since these metrics are the most used evaluation metrics in machine learning (Naidu et al., 2023). However, Naidu et al. (2023) mentions that these basic metrics should not be used in isolation. I will, therefore, also consider the ROC and AUC metrics to evaluate the model's sensitivity to precision and recall. The evaluation metrics of these models are compared to those of a simple dummy classifier, serving as the baseline.

Furthermore, permutation feature importance is computed on the test set to get insight into which features are providing high predictive accuracy. It shows whether the proxies for the availability bias, as well as the interaction terms between the proxies for the availability bias and the number of lives saved, are "important" for the models predictive power. Permutation feature importance evaluates how much the model's prediction error increases when the values of a specific feature are randomly

shuffled (Molnar, 2020). As noted by Molnar (2020), a feature is considered "important" if shuffling its values leads to an increase in model error, indicating the model depends on that feature for its predictions. A feature is deemed "unimportant" if shuffling its values has no impact on the model error, suggesting the model did not use the feature in its predictions. The method was first introduced by Breiman (2001), who used it on a random forest model. Since then, it have been used on many different models including logistic regression and neural networks (Potter, 2005; J.-B. Yang et al., 2009).

Partial dependence plots are then used to visualize and evaluate the interaction effect of the two most important interaction terms identified by permutation feature importance. These plots make it easier to understand the relationship between the input features and the target feature, especially for complex models such as random forest and neural network (Greenwell et al., 2017).

4.3.4 *Binary Logistic Regression*

Binary logistic regression is a statistical method used to predict a binary target variable with two categories based on one or more independent variables, which can be continuous or categorical (Sreejesh et al., 2014). This paper applies a binary logistic regression model as it has been used before in the context of AV crashes by Houseal et al. (2022). The model estimates coefficients for each independent variable by maximizing the likelihood that the observed data match the predicted probabilities (Sreejesh et al., 2014). Coefficients indicate the direction and magnitude of the relationship between the independent variables and the log odds of the dependent variable. The logistic function converts the log odds into a probability for one class (Sreejesh et al., 2014).

The model is implemented in Python using the LogisticRegression function from Scikit-learn. To optimize its performance, a random search is employed to find the best parameters. the search space includes various regularization functions and strengths to apply a penalty to the error term. Additionally, the optimization algorithm and the maximum number of iterations for convergence are selected randomly. Different stopping criteria are tested, meaning the solver stops if the change in the cost function between iterations is smaller than a specified threshold. The hyper-parameter values are shown in Table 1

Hyper-parameter	Values
Regularization (penalty)	['l1', 'l2', 'elasticnet', None]
Regularization strength (C)	[0.001, 0.01, 0.1, 1, 10, 100]
Optimization algorithm (solver)	['liblinear', 'saga', 'lbfgs']
Maximum iterations	[100, 200, 500]
Stopping criteria (tol)	[0.0001, 0.001, 0.01]
Class Weights	[None, 'balanced']

Table 1: Hyper-parameters and their corresponding values for the binary logistic regression model.

4.3.5 *Decision Tree*

A simple decision tree model will also be employed in this paper, as it is one of the most widely used methods in traffic crash data analysis and can capture non-linear relationships (Ashraf et al., 2021; Houseal et al., 2022). According to Song and Ying (2015), a decision tree is a hierarchical model used for classification, consisting of root, internal, and leaf nodes connected by branches. It makes decisions through "if-then" rules, where each path from root to leaf represents a classification outcome (Song and Ying, 2015). The tree is built through splitting, which divides nodes using the most important input features, related to the target variable. This process continues until stopping criteria are met.

The model is implemented in Python using the DecisionTreeClassifier function from Scikit-learn. The optimal set of parameters is found by testing for different combinations of the maximum depth of the tree, the minimum number of samples to split an internal node, the minimum number of samples that must be present in a leaf node and the quality measure of a split. The values of the hyper-parameters are shown in Table 2

Hyper-parameter	Values
Splitting criterion	['gini', 'entropy']
Maximum depth	[10, 15, 20, 25, None]
Minimum sample split	[10, 20, 30, 40]
Minimum sample leaf	[5, 10, 20, 30]

Table 2: Hyper-parameters and their corresponding values for the decision tree model.

4.3.6 *Random Forest*

Random forests consist of multiple decision trees, where each tree is built using a random subset of the data (Breiman, 2001). The trees are created independently, and the same probability distribution is used to generate

the random data for each tree in the forest. According to Breiman (2001), growing multiple trees, as in a random forest, can significantly enhance classification accuracy by combining the predictions of multiple decision trees through a voting mechanism. This mechanism reduces the risk of over-fitting and improves generalization compared to relying on a single decision tree (Breiman, 2001).

The model is implemented in Python using the RandomForestClassifier function from Scikit-learn. As at the decision tree model, the hyper-parameters tuned at the random forest model also include the maximum depth of the tree, the minimum number of samples to split an internal node, the minimum number of samples that must be present in a leaf node and the quality measure of a split. Additionally, the model performance is evaluated on a different number of trees in the forest. The values of the hyper-parameters are shown in Table 3

Hyper-parameter	**Values**
Splitting criterion	['gini', 'entropy']
Maximum depth	[10, 15, 20, 25, None]
Minimum sample split	[20, 30, 40, 50]
Minimum sample leaf	[20, 30, 40, 50]
Number of trees	[100, 150, 200, 250, 300]

Table 3: Hyper-parameters and their corresponding values for the random forest model.

4.3.7 *Neural Network*

Neural networks will be used to model more complex interactions and patterns within the data, potentially identifying patterns that the previous models might miss. Neural networks were first introduced by McCulloch and Pitts (1943). Murtagh (1991) describes a neural network as a set of layers of units, called neurons, with connections between all neurons in one layer and all neurons in the next layer. Neural networks consist of an input layer, one or more hidden layers, and an output layer (Murtagh, 1991). The input from one layer goes through an activation function of which the output is passed to the next layer, with the final output produced by the outermost layer (Sharma et al., 2017). Neural networks learn by a process called back-propagation, first described by Rumelhart et al. (1986). The procedure repeatedly adjusts the connection weights based on the error between the predicted and the actual outputs. Deep learning techniques are well-suited for large datasets, making this approach appropriate for this research (Mahesh, 2020). Moreover, neural networks have alreeady

been used to predict human decisions in AV accidents by Wiedeman et al. (2020) and Agrawal et al. (2019)).

The neural network is implemented in Python using TensorFlow Keras, with components from the Keras library for model architecture, layers, and optimizers. The Scikeras Wrappers package is used to wrap the Keras model as a Scikit-learn classifier. Several model architectures are trained, the model architecture from Wiedeman et al. (2020) and 25 random architectures through random search.

The model architecture by Wiedeman et al. (2020) is shown in Figure 6. It consists of two densely connected hidden layers with both 64 neurons, batch normalization, a ReLU activation function, and an output layer with a sigmoid activation function. Binary cross entropy is used as the loss function, and the optimizer is RMSprop (Wiedeman et al., 2020). The number of input neurons differs from the 24 in Figure 6 because this paper feeds 39 features to the network instead of 24.

The hyper-parameters varied during random search include the number of hidden layers, the number of neurons per layer, the activation function, the optimizer, and the number of training epochs. The values of these hyper-parameters are shown in Table 4.

Hyper-parameter	**Values**
Number of hidden layers	[1, 2, 3, 4]
Number of Neurons	[16, 32, 64, 128]
Activation functions	['relu', 'leaky_relu', 'tanh']
Epochs	[10, 20, 30, 40, 50]
Optimizers	['adam', 'rmsprop']

Table 4: Hyper-parameters and their corresponding values for the neural network.

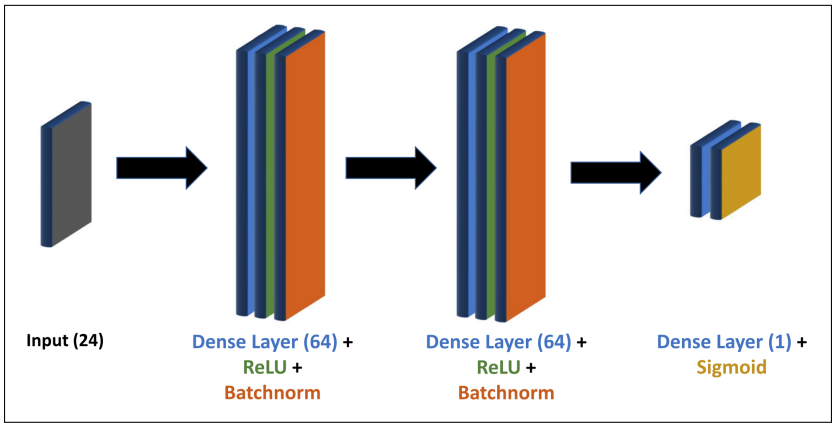


Figure 6: Graphical presentation of the neural network architecture by Wiedeman et al. (2020)

5 RESULTS

This section reports the results of this study. The learning algorithms are compared, the feature importance for the features in the "complete" dataset and the interaction terms in the "utilitarian" dataset are discussed and partial dependence plots are shown.

5.1 *Model Comparison (SQ1)*

The first sub-question investigates the performance of the machine learning models binary logistic regression, decision tree, random forest, and neural networks when predicting moral decisions in the Moral Machine scenarios. First, we compare model scores. Later, in subsection 5.1.1, we test the best model on the test set to understand the implications of these scores in the context of road traffic scenarios in the Moral Machine.

Table 5 shows the results of the best models from each algorithm. Table 7 in Appendix C lists the parameters of each best performing model. Overall, the models perform similarly, with only small differences between them. The dummy classifier has a score of 0.500 for each metric, as expected for an untrained baseline, predicting randomly with equal probability. Given an evenly distributed binary target variable, each metric is 0.500.

Random forest performs as one of the best models across most metrics, including an accuracy of 71.4%. This indicates that the random forest model correctly predicts the target variable whether people want to save or kill characters in the Moral Machine in 71.4% of the cases. The model consists of 250 trees with a maximum depth of 25 branches, a minimum of 50 samples required to split a node, and a minimum of 20 samples required per leaf node. Entropy is used to measure the quality of the split. Other models show similar performance to random forest. The best neural network model during random search and the one by Wiedeman et al. (2020) have nearly identical scores. As shown in Table 7, the best neural network consists of three hidden layers with batch normalization, a tanh activation function, and a sigmoid output layer. It uses 32 neurons per layer, 50 training epochs, and RMSprop optimizer. This neural network has slightly lower precision (0.719) but higher recall (0.702) and AUC (0.781) compared to random forest. It has the same accuracy but a slightly better balance between precision and recall, making it the most effective model overall. The decision tree achieves good precision (0.724) but falls short on recall (0.679) and F1-score (0.700), indicating it struggles with balanced predictions. Binary logistic regression has the lowest scores for precision (0.714), AUC (0.765), and accuracy (0.707), making it less effective than more complex models.

In summary, the neural network appears to be the best model overall, but the differences between the neural network and random forest are small. The scores indicate that all the models (except the baseline) are capable of providing reasonable predictions for this task.

	Evaluation Metric				
Model	**Precision**	**Recall**	**F1-Score**	**AUC**	**Accuracy**
Dummy classifier	0.500	0.500	0.500	0.500	0.500
Logistic regression	0.714	0.691	0.702	0.765	0.707
Decision tree	0.724	0.679	0.700	0.775	0.710
Random forest	0.721	0.698	0.710	0.780	0.714
Wiedeman et al. (2020)*	0.718	0.704	0.711	0.780	0.713
Neural network	**0.719**	**0.702**	**0.710**	**0.781**	**0.714**

Table 5: Model performances for binary predictions whether people save or kill characters in road traffic accidents with AVs ("complete" dataset). *Neural network architecture by Wiedeman et al. (2020).

5.1.1 *Best Model Performance*

Since the neural network is the best performing model, it was tested on a 20% hold-out test set. It achieved balanced performance across metrics, with accuracy, precision, recall, and F1-score all at 0.714, suggesting consistent classification capabilities. This means that the network correctly predicts 71.4% of decisions in the Moral Machine. From the outcomes where people chose to save characters, 71.4% were correctly classified, and in 71.4% of cases where the network predicted participants saving characters, participants actually made that choice. The AUC score is equal to 0.780, indicating that the neural network assigns a random outcome where people wanted to save characters a higher probability of being 'saved' than a random outcome where people wanted to kill characters in 78.0% of the cases.

The confusion matrix in Figure 7 shows the model correctly classified 98,625 'save' outcomes and 102,122 'kill' outcomes. However, there is a notable number of misclassifications, with 38,402 false positives and 41,898 false negatives. This means that many 'save' outcomes were labeled as 'kill' and vice versa.

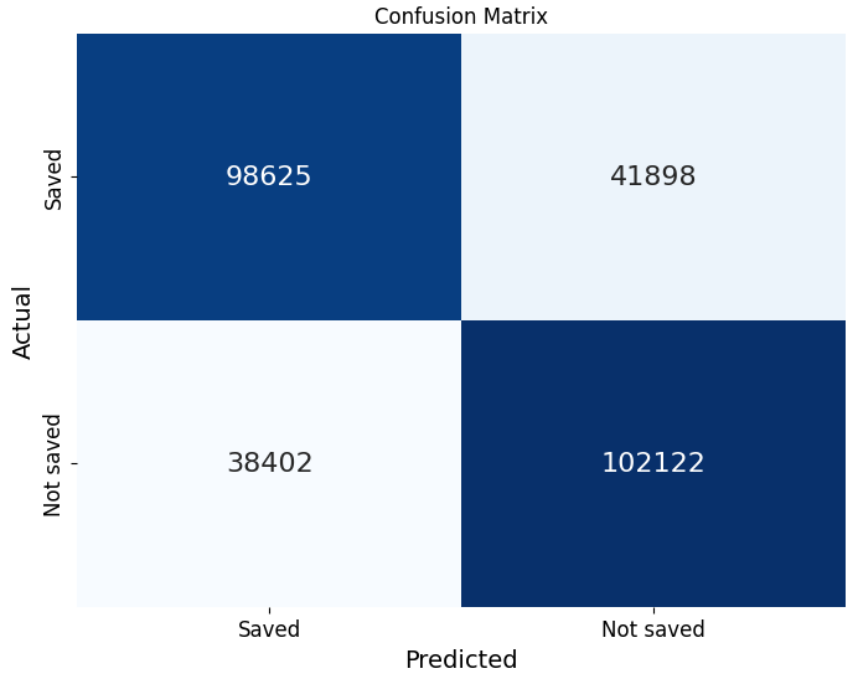


Figure 7: Confusion matrix for the "complete" test set.

The ROC Curve in Figure 8 illustrates the trade-off between true positive rate and false positive rate as the classification threshold changes (Hoo et al., 2017). The true positive rate measures the model's ability to correctly predict outcomes where people wanted to save characters. The false positive rate measures how often the model incorrectly predicts outcomes where people want to kill characters. The different thresholds are represented along the curve, where for each threshold, the true positive and false positive rate are shown. Lower thresholds lead to more 'save' predictions, increasing true positive rates but also false positive rates. Higher thresholds reduce 'save' predictions, decreasing false positive rates but also true positive rates. Ideally, the curve should touch the top-left corner, representing a high true positive rates and low false positive rates (Hoo et al., 2017). The curve lies well above the diagonal, showing that the model performs better than random guessing when predicting human moral decisions. However, there may still be significant false positives or false negatives.

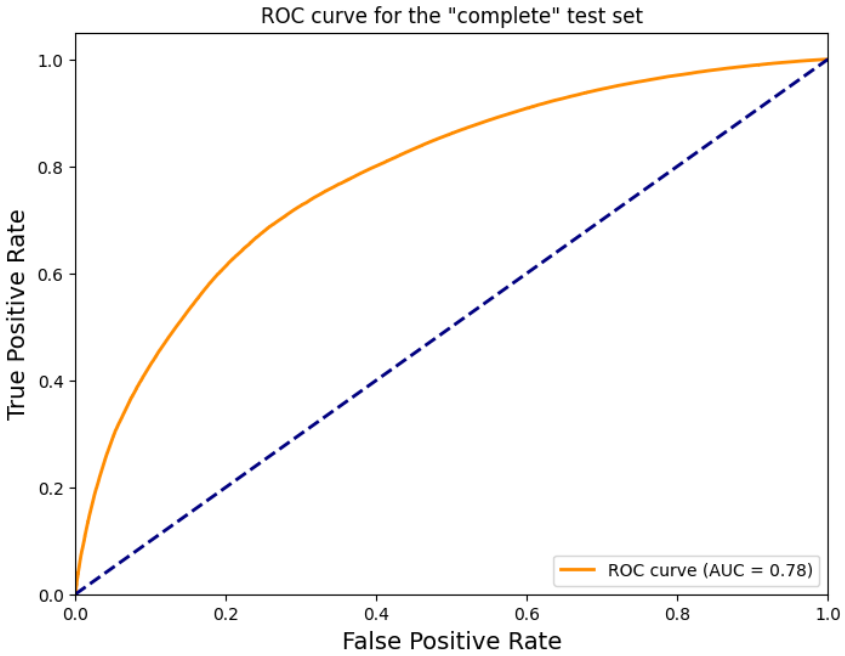


Figure 8: ROC curve for the "complete" test set.

The neural network's training and test scores are almost identical, indicating that it generalizes well without overfitting. The model learned the patterns in the training data effectively and applies them to new data, but there is room to reduce misclassifications and improve metrics.

5.2 *Feature Importance (SQ2)*

The second sub-question investigates whether the estimated number of road traffic deaths for males, females, pedestrians, and passengers, along with statistics on the frontier technology readiness index, predict moral decisions in the Moral Machine scenarios. Permutation feature importance measures the contribution of each feature to the model's performance.

The barplot in Figure 9 shows the average importance of features in different feature groups for predicting decisions in the Moral Machine experiment. 'Scenario' features describe the outcome, such as the number and type of characters or the presence of a traffic light. 'Road Traffic Death' features include statistics on traffic deaths, while 'Technology Readiness Index' features include statistics such as 'ICT', 'Skills', 'Research and Development', 'Industry activity', 'Access to finance', and an 'Overall index'.

Features describing the scenario of an outcome contribute by far the most to the model's decision-making process. On average, 2.2% of the

total feature importance in the neural network model is attributed to each feature in the "scenario features" group. This means the model relies heavily on scenario-specific features to make predictions.

On average, each statistic on road traffic deaths accounts for 0.6% of the model's predictive power, while each Technology Readiness Index statistic accounts for 0.2%. Thus, features serving as proxies for the availability bias do influence the network decisions, but to a much lesser degree than scenario features.

The K-Prototype Clustering features show minimal importance, contributing only a small fraction to overall predictive power. This indicates that the clustering features from K-Prototype clustering do not significantly improve the model's ability to predict outcomes.

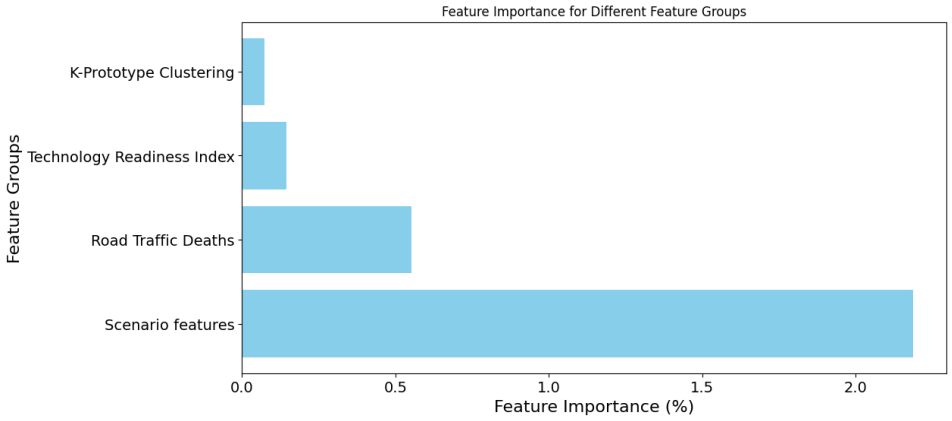


Figure 9: Feature importance for different feature groups

The specific feature importance of the features serving as proxy for the availability bias are shown in Figure 10. The estimated number of road traffic deaths for the total population in a county, and for males and females separately, seem to contribute the most to the models predictive power. It is followed by a lower feature importance for the overall technology readiness index. Prior road traffic accidents in general, accidents including male and female, and a country's readiness for frontier technologies such as AVs, are important indicators of the decisions people want AVs to make. However, only the number of road traffic deaths has a feature importance higher than 1%, the rest of the feature importances are still only less then 0.8% compared to the average of 2.2% that the "scenario" features contribute to the models predictive power. Other features serving as proxy for the availability bias, such as the level of industry activity, the level of research and development and the level of ICT infastructure, only contribute less than 0.3% to the model's predictive power. A country's access to finance, the estimated number of road traffic deaths for passengers and the level

of relevant skills are the least important for the model's ability to predict decision outcomes.

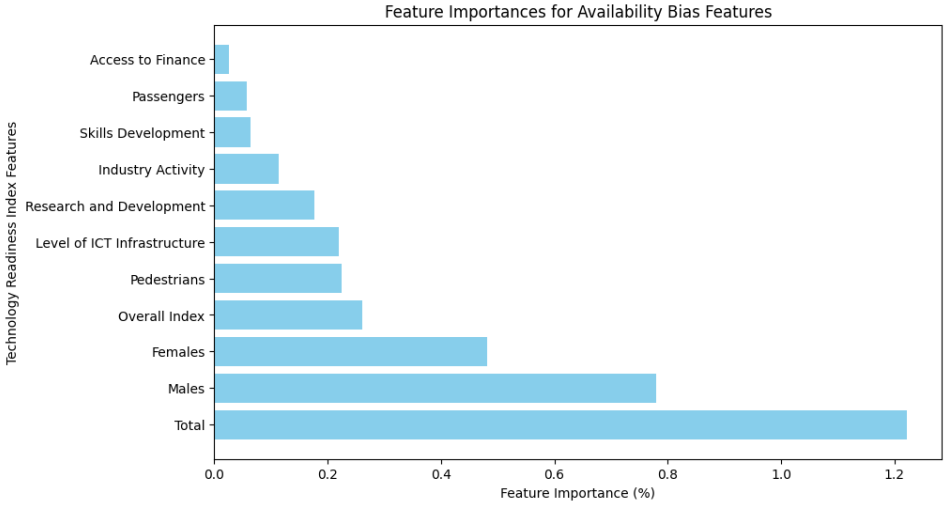


Figure 10: Feature importance for the availability bias features

5.3 *Utilitarian Subset (SQ3)*

The third sub-question asks whether the preference for utilitarian AVs is affected by the availability bias. To answer this question, we show how the effect of the number of lives that can be saved in an outcome on the probability of people saving the characters in that outcome is dependent on prior knowledge about AVs and past road traffic accidents. An additional dataset is created only including "utilitarian" scenarios. In these scenarios, the characters are the same on both sides, in addition to at least one more character on one side. These dataset includes a variable indicating the number of lives saved in an outcome and 11 interaction features between the proxies for the availability bias and the number of lives saved.

5.3.1 *Best Model Performance*

Table 6 shows the results of the best performing models of the different learning algorithms trained on the "utilitarian" dataset.

As at the model performances for the "complete" dataset, the results indicate that all models perform similarly, making it difficult to distinguish a clear preference between them. However, we can see that their performance is significantly better than the baseline model, the dummy classifier. Moreover, the performances of all models significantly surpass the performances achieved on the "complete" dataset including all scenario types (Table 5). While the evaluation metrics of all the models trained on

the "complete" dataset only slightly surpass 0.7, the evaluation metrics of the model trained on the "utilitarian" dataset are all higher than 0.77. When training the best performing model from the random search on the test set, we get evaluation metrics all equal to 0.775. This indicates that it is easier for the learning algorithms to predict whether people want an AV to save or kill individuals in road traffic accidents if the number of individuals saved or killed in an outcome differs from the alternative outcome.

As can be seen in the next two subsection, subsection 5.3.2 and 5.3.3, the better model performance on the "utilitarian" dataset is indeed due to the variable indicating the number of lives saved in an outcome. It is also due to the interaction terms included in the "utilitarian" dataset and due to the fact that the effect of the number of lives that can be saved in an outcome on the decision in that outcome depends on prior knowledge about AVs and prior road traffic accidents.

	Evaluation Metric				
Model	**Precision**	**Recall**	**F1-Score**	**AUC**	**Accuracy**
Dummy classifier	0.500	0.500	0.500	0.500	0.500
Logistic regression	0.771	0.779	0.775	0.827	0.774
Decision tree	0.771	0.778	0.775	0.775	0.774
Random forest	0.772	0.778	0.774	0.839	0.774
Wiedeman et al. (2020)*	0.771	0.777	0.774	0.839	0.773
Neural network	**0.773**	**0.778**	**0.775**	**0.840**	**0.774**

Table 6: Model performances for binary predictions whether people save or kill characters in road traffic accidents with AVs ("utilitarian" dataset). *Neural network architecture by Wiedeman et al. (2020).

5.3.2 *Interaction terms*

The blue bars in Figure 11 show the average importance of features in different feature groups for the "complete" dataset, just as in Figure 9. The red bars represent the average importance of features in different groups for the "utilitarian" dataset. The "Road Traffic Deaths" features for the "utilitarian" dataset include interaction terms between the number of lives saved and the number of road traffic deaths for males, females, pedestrians, and passengers. The "Technology Readiness Index" features include the importance of interaction terms between the number of lives saved and frontier technology readiness index statistics.

In Figure 11, the red bars for the "Technology Readiness Index" and "Road Traffic Deaths" feature groups are significantly higher than the blue bars. This indicates that the interaction terms between the number of lives saved and road traffic deaths, as well as the frontier technology

readiness index, significantly increase the average importance of these feature groups. These interaction terms improve the predictive power of the models. Interaction terms between "Number of lives saved" and features related to prior information about AVs thus have a higher feature importance than the features alone.

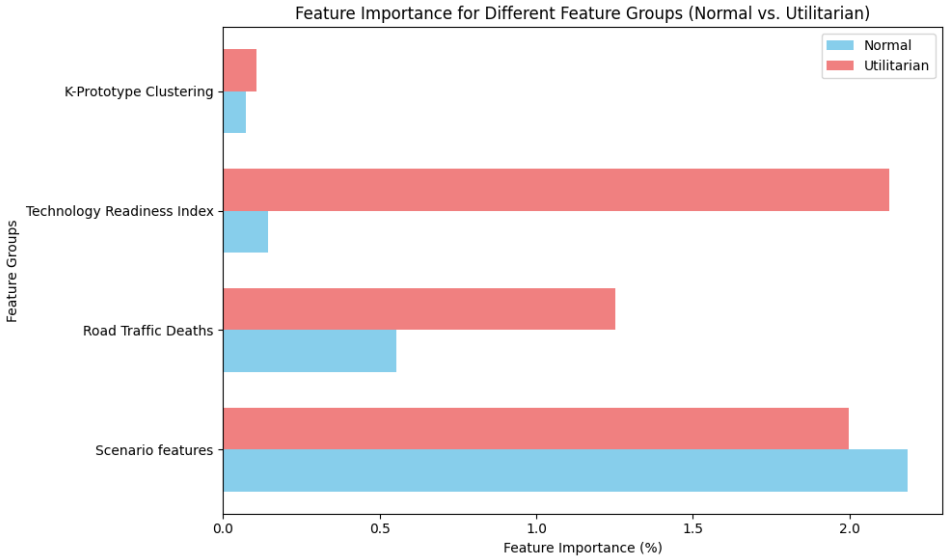


Figure 11: Feature importance for different feature groups including interaction terms

Figure 12 shows the permutation feature importance for all features in the "utilitarian" dataset. It can be seen that the feature "Number of Lives Saved" (yellow) has a relatively high feature importance. The number of lives that can be saved in an outcome are thus really important for the model's ability to predict decision outcomes. As we could already conclude from Figure 11, the plot highlights that most of the interaction terms (red) also have a relatively high feature importance. The interaction term "Skills level" has the highest feature importance and ranks significantly higher than its standalone feature (blue). The interaction term "Passengers" also has a significantly higher feature importance than its standalone feature. This suggests that the neural network captures the more complex relationships between features and that the combined effect of these interactions provides greater predictive power than the individual features alone.

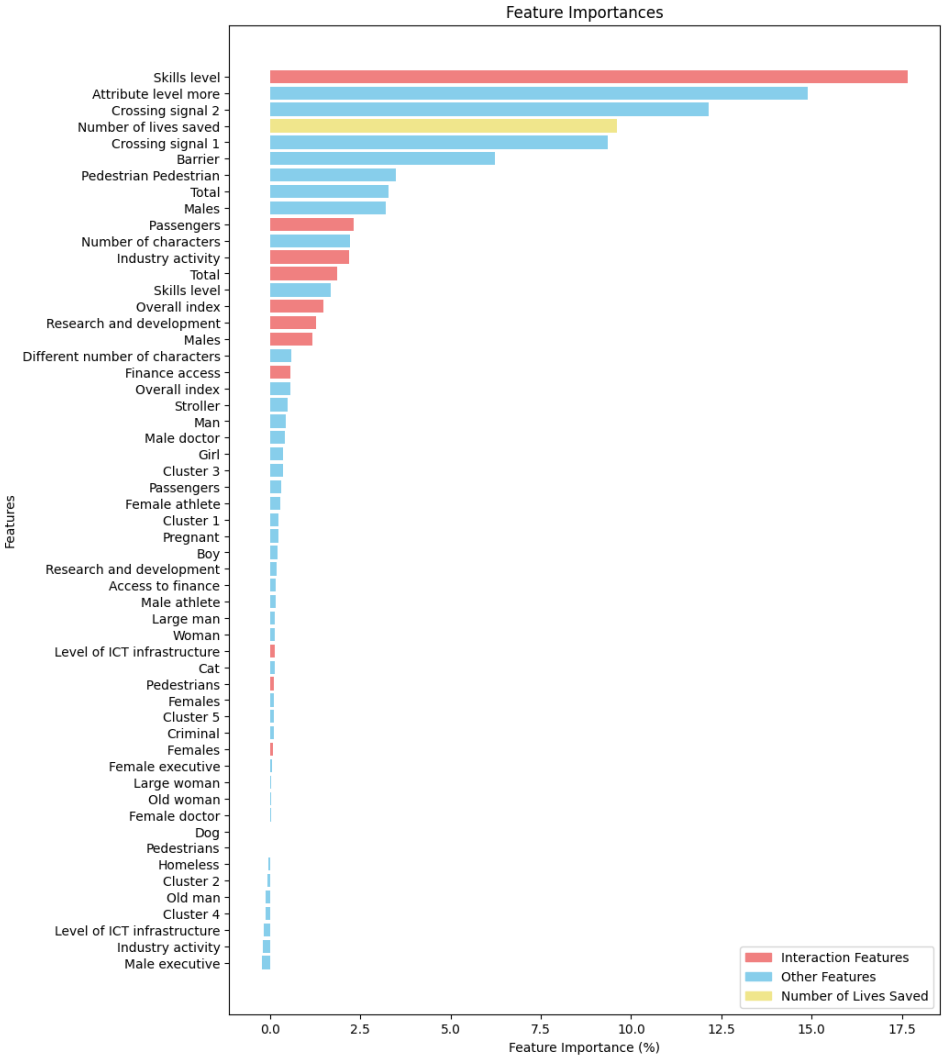


Figure 12: Feature importance for all features in the "utilitarian" dataset

5.3.3 *Partial Dependence Plot*

The interaction effect of the two most important interaction terms in Figure 12, "Skills level" and "Passengers", are shown in the partial dependence plot in Figure 13 and Figure 14. The figures show how the effect of the number of lives that can be saved in a scenario on the probability of people saving the characters in that scenario dependents on prior knowledge about AVs and past road traffic accidents.

Figure 13 shows how the effect of the number of lives that can be saved in a scenario on the probability of people saving characters in that scenario depends on the level of relevant skills for using, adopting and adapting frontier technologies. First, it can be seen that as the skills level increases,

the probability of people saving characters decreases in general. This suggests that higher skill levels are associated with a smaller likelihood of people saving characters. The plot also shows a clear interaction term. For low skill levels, the probability of people saving characters increases as the number of lives that can be saved in a scenario increases. However, for high skill levels, the probability of people saving characters decreases as the number of lives that can be saved in a scenario increases. This means that people with a higher skill level are less likely to endorse utilitarian AVs programmed to save the largest number of road users, compared to people with a lower skill level.

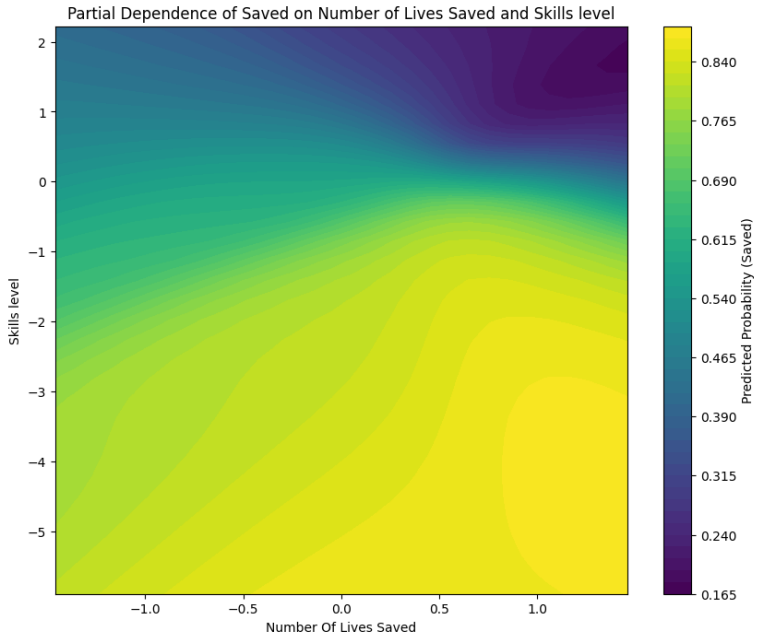


Figure 13: Partial dependence plot of probability of choosing save and the interaction of the number of lives saved and the skills level.

Figure 13 shows how the effect of the number of lives that can be saved in a scenario on the probability of people saving characters in that scenario depends on the estimated number of road traffic deaths for passengers. First, it can be seen that for a higher number of passenger deaths, the probability of people saving characters is lower in general. Only for the lowest number of passenger deaths, the probability of people saving characters is higher. This suggests that a low number of passenger deaths is associated with a larger likelihood of people saving characters. This plot also shows a clear interaction term. For a low number of passenger deaths, the probability of people saving characters increases as the number of lives that can be saved in a scenario increases. However, for a moderate and high

number of passenger deaths, the probability of people saving characters decreases as the number of lives that can be saved in a scenario increases. This means that people more often exposed to road traffic accidents with passengers are less likely to endorse utilitarian AVs and thus more likely to save less number of lives.

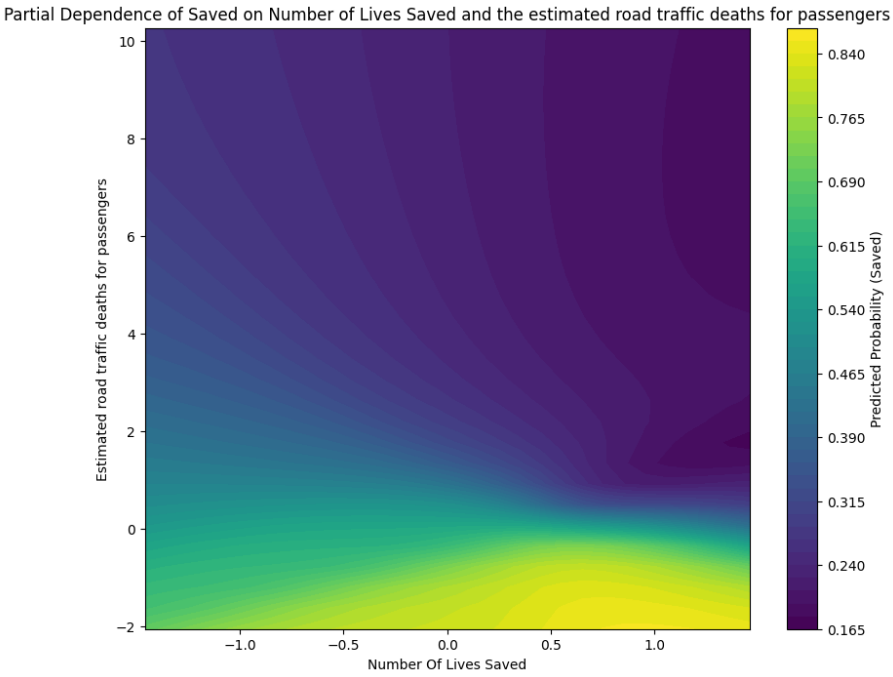


Figure 14: Partial dependence plot of probability of choosing save and the interaction of the number of lives saved and the estimated number of road traffic deaths for passengers.

6 DISCUSSION

The goal of this research was to predict human moral decisions regarding whether to save or kill individuals in fatal road traffic accidents involving AVs. The focus of this paper was on the effect of prior knowledge about AVs and prior road traffic accidents on human moral decisions. Various machine- and deep-learning models were compared to predict these decisions. Additionally, this paper looked if the effect of the number of lives that can be saved in a scenario on the decision of people in that scenario is dependent on prior knowledge about AVs and past road traffic accidents.

SQ1 - Using features describing the characteristics of a specific outcome, features on prior information about AVs, clustering techniques and various different learning algorithms, we were able to obtain an accuracy score of

71.4% on the test set. This means that the best model correctly predicts the target variable whether people want to save or kill characters in the Moral Machine in 71.4% of the cases. Creating a dataset with only one specific scenario type and eleven additional interaction terms resulted in 77.4% accuracy. Both accuracies fall within the range of those reported in previous studies predicting decisions in the Moral Machine. Prior research, utilizing neural networks, namely achieved accuracies ranging from 66.0% to 77.4%. The best performing model on the validation folds in the present paper was also a neural network. However, the Random forest performed equally well and the differences in scores across the other models are relatively small. While the neural network and random forest model thus perform better than the more simple decision tree and logistic regression model, all models are capable of providing reasonable predictions for this task. This is in contrast with the findings from Wiedeman et al. (2020) and Agrawal et al. (2019), who demonstrated that deep learning models significantly outperform simpler models in understanding moral decisions in the Moral Machine dataset. **SQ2 -** Both Zhu et al. (2022) and Othman (2023) found that prior information about AVs significantly influenced people's moral judgments. However, the results derived from permutation feature importance in the present paper cannot support these findings. The estimated number of road traffic deaths for different groups and the frontier technology readiness index did not have a significant contribution to the model's predictions. The feature importance of these features was significantly lower than the features describing the characteristics of the scenario. Prior information about AVs did thus not impact peoples preference to save or kill characters in the Moral Machine. **SQ3 -** It did, however, affect the preference for utilitarian AVs. This paper demonstrates, through feature engineering interaction terms and the use of partial dependence plots, that people with higher skill levels are less likely to endorse utilitarian AVs, compared to those with lower skill levels. This aligns with the findings by Zhu et al. (2022), who suggest that people in the AV industry, and thus more expertise and relevant skills, were less likely to endorse utilitarian algorithms compared to those outside of the AV industry. The findings do however contradict with Othman (2023), who showed that people with prior knowledge about AVs were more likely to endorse utilitarian algorithms in AVs. The present paper also demonstrates that people more often exposed to road traffic accidents with passengers are less likely to endorse utilitarian AVs and thus more likely to save less number of lives. This corresponds to the findings by Othman (2023), who showed a significant decrease in people favoring utilitarian algorithms after they were exposed to traffic accidents.

6.1 *Limitations*

One limitation of this study is the use of regular road traffic accident data instead of crash data specifically involving AVs. This was due to the limited availability of AV crash databases (Ashraf et al., 2021), making it challenging to study AV-related decision-making. Consequently, the findings might not fully reflect the impact of past AV crashes on ethical decisions.

Additionally, only a small subset of the dataset was used for the Elbow method in K-prototype clustering. Running clustering algorithms on large datasets can be computationally expensive and time-consuming (Nerurkar et al., 2018). Using a smaller subset allowed for an efficient approximation of the optimal number of clusters without excessive computation time. Once identified, this number was applied to the entire dataset.

The main limitation is the small difference in performance between the learning algorithms. The neural network did not significantly outperform the simpler models. As shown in the learning curves in Figure 17 and Figure 18 in Appendix F, both the training and validation accuracy for the "complete" and "utilitarian" datasets increase in the frist epochs and then plateau. This might indicate insufficient model capacity, suggesting the need for more layers, neurons, or increased complexity. Alternatively, it may suggest that the current features or data lack the necessary information for further improvements. However, similar simple networks have been proven to be effective in related work (Wiedeman et al. (2020), Agrawal et al., 2019) in the context of the Moral Machine. Furtheromore, random search was employed to explore different architectures, with normalized and encoded data, and additional engineered features. Future improvements could involve analyzing the gradient during training to diagnose issues like the vanishing gradient problem and ensuring that the network updates its weights as expected. The vanishing gradient problem refers to the behaviour where the weights in a neural network go exponentially fast to 0 through backpropoagation, making it impossible for the model to learn (Pascanu et al., 2012). This issue could explain the plateau observed in Figure 17 and Figure 18.

7 CONCLUSION

RQ - The main research question addressed in this paper was whether the availability bias significantly influences human moral decisions in scenarios involving AVs. The findings presented in this paper demonstrate the ability to predict decisions made in the Moral Machine experiment, but no evidence was found to suggest that the availability bias directly influenced

moral decisions. Prior knowledge about AVs and past experiences with road traffic accidents did not have a direct impact on human decision-making. However, it did affect decisions indirectly, interacting with the number of lives that could be saved in AV crashes. The paper found that people with higher skill levels were less likely to endorse utilitarian AVs compared to those with lower skill levels. Additionally, people who had been exposed to road traffic accidents involving passengers were less likely to favor utilitarian AVs.

These findings suggest that actions could be taken to promote utilitarian perspectives, often regarded as the morally responsible choice when considering the perspective of all road users involved in an accident (Othman, 2023). In countries with a low preference for utilitarian AVs, either due to a high number of road traffic accidents or a high skill level, efforts should be made to increase overall utilitarian judgments (Othman, 2023). Moreover, designers of AVs often have high skill levels and therefore a lower preference for utilitarian AVs. This low preference for algorithm may then be inscribed into technology by the designers of the AVs (Poszler et al., 2023). Regulations or guidelines may be necessary to prevent this form happening and to influence how AV algorithms are programmed.

The findings of this paper contribute to the understanding of human decision-making and how biases may influence the design and functioning of AI algorithms in AVs. This knowledge can inform targeted actions to promote utilitarian AVs. These efforts can help address the challenge of designing a fair and consistent ethical framework for AV algorithms (Frank et al., 2019).

REFERENCES

Abadi, M., Agarwal, A., Barham, P., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org].

Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M., & Gardner, J. L. (2016). Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences*, *113*(25), E3548–E3557.

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*.

Alam, M. (2020). Data normalization in machine learning. *Towards Data Science. December*.

Ashraf, M. T., Dey, K., Mishra, S., & Rahman, M. T. (2021). Extracting rules from autonomous-vehicle-involved crashes by applying decision

tree and association rule methods. *Transportation research record*, *2675*(11), 522–533.

Awad, E. (2021, "Oct"). Moral machine. https://osf.io/3hvt2/

Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, *63*(3), 48–55.

Awad, E., Dsouza, S., Chang, P., Rahwan, I., Bonnefon, J.-F., Shariff, A., & Tang, D. (2024). Moral machine [Accessed: 2024-11-27]. https://www.moralmachine.net/

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in managerial decision making*. John Wiley & Sons.

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. *Frontiers in behavioral neuroscience*, *12*, 31.

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

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Conference on fairness, accountability and transparency*, 149–159.

Bodenschatz, A. (2024). When own interest stands against the "greater good"–decision randomization in ethical dilemmas of autonomous systems that involve their user's self-interest. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100097.

Bodenschatz, A., Uhl, M., & Walkowitz, G. (2021). Autonomous systems in ethical dilemmas: Attitudes toward randomization. *Computers in Human Behavior Reports*, *4*, 100145.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

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

Chira, I., Adams, M., & Thornton, B. (2008). Behavioral bias within the decision making process.

Chollet, F., et al. (2015). Keras.

Combs, B., & Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism quarterly*, *56*(4), 837–849.

DATA-PERSON. (n.d.). Scikeras: Scikit-learn api keras.

Dattner, B., Chamorro-Premuzic, T., Buchband, R., & Schettler, L. (2019). The legal and ethical implications of using ai in hiring. *Harvard Business Review*, *25*, 1–7.

Ehmann, A. (2023). Pycountry: Iso country, subdivision, language, currency and script definitions.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and engineering ethics*, *25*, 399–418.

Foundation, P. S. (2024). Python [Accessed: 2024-11-29].

Frank, D.-A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific reports*, *9*(1), 13080.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, *17*(3), 295–314.

Gong, J., & Chen, T. (2022). Does configuration encoding matter in learning software performance? an empirical study on encoding schemes. *Proceedings of the 19th International Conference on Mining Software Repositories*, 482–494.

Google. (2024). Google colab [Accessed: 2024-11-29].

Greenwell, B. M., et al. (2017). Pdp: An r package for constructing partial dependence plots. *R J.*, *9*(1), 421.

Habla, W., Kataria, M., Martinsson, P., & Roeder, K. (2024). Should it stay, or swerve? trading off lives in dilemma situations involving autonomous cars. *Health economics*, *33*(5), 929–951.

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

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an roc curve?

Houseal, L. A., Gaweesh, S. M., Dadvar, S., & Ahmed, M. M. (2022). Causes and effects of autonomous vehicle field test crashes and disengagements using exploratory factor analysis, binary logistic regression, and decision trees. *Transportation research record*, *2676*(8), 571–586.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values," proceedings of 1st pacific-asia conference on knowledge discouvery and data mining.

Hug, N. (2017). Kmodes: Python implementations of k-modes and k-prototypes clustering.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

International Labour Organization. (2024). International labour organization (ilo). https://www.ilo.org/

International Monetary Fund. (2024). International monetary fund (imf) - home page. https://www.imf.org/en/Home

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

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, *93*(5), 1449–1475.

Kai-Ineman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J. B., & Rahwan, I. (2018). A computational model of commonsense moral decision making. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 197–203.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of experimental psychology: Human learning and memory*, *4*(6), 551.

Lim, H. S. M., & Taeihagh, A. (2019). Algorithmic decision-making in avs: Understanding ethical and technical concerns for smart cities. *Sustainability*, *11*(20), 5791.

Liu, P., & Liu, J. (2021). Selfish or utilitarian automated vehicles? deontological evaluation and public acceptance. *International Journal of Human–Computer Interaction*, *37*(13), 1231–1242.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS one*, *16*(12), e0261673.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115–133.

McKinney, W. (2011). Pandas: A foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*, 1–9.

Microsoft. (2024). Visual studio code [Accessed: 2024-11-29].

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5-6), 183–197.

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. *Computer Science On-line Conference*, 15–25.

Nerurkar, P., Shirke, A., Chandane, M., & Bhirud, S. (2018). Empirical analysis of data clustering algorithms. *Procedia Computer Science*, *125*, 770–779.

Ng, Y.-L. (2024). Understanding passenger acceptance of autonomous vehicles through the prism of the trolley dilemma. *International Journal of Human–Computer Interaction*, *40*(9), 2185–2194.

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A voting-based system for ethical decision making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

OpenAI. (2024). Chatgpt (version 4) [Accessed: 2024-11-29].

Organisation for Economic Co-operation and Development. (2024). Oecd - home page. https://www.oecd.org/

Othman, K. (2023). Understanding how moral decisions are affected by accidents of autonomous vehicles, prior knowledge, and perspective-taking: A continental analysis of a global survey. *AI and Ethics*, 1–18.

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). On the difficulty of training recurrent neural networks. *30th International Conference on Machine Learning, ICML 2013*.

PatSeer. (2024). Patseer - patent search and analysis. https://patseer.com/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poszler, F., Geisslinger, M., Betz, J., & Lütge, C. (2023). Applying ethical theories to the decision-making of self-driving vehicles: A systematic review and integration of the literature. *Technology in Society*, 102350.

Potter, D. M. (2005). A permutation test for inference in logistic regression with small-and moderate-sized data sets. *Statistics in medicine*, *24*(5), 693–708.

Rich, A., & Gureckis, T. (2019). Lessons for artificial intelligence from the study of natural stupidity. nat mach intell 2019; 1: 174–80. https://doi.org/10.1038/s42256-019-0038-z

Ronacher, A. (2023). Jinja2 - templating engine for python.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, *61*(2), 195.

Scopus. (2024). Scopus search page. https://www.scopus.com/search/form.uri?display=basic#basic

Shaheen, M. Y. (2021). Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*.

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696.

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, *6*(12), 310–316.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Sreejesh, S., Mohapatra, S., Anusree, M., Sreejesh, S., Mohapatra, S., & Anusree, M. (2014). Binary logistic regression. *Business Research Methods: An Applied Orientation*, 245–258.

Sui, T. (2023). Exploring moral algorithm preferences in autonomous vehicle dilemmas: An empirical study. *Frontiers in Psychology*, *14*, 1229245.

Sushina, T., & Sobenin, A. (2020). Artificial intelligence in the criminal justice system: Leading trends and possibilities. *Proceedings of the 6th International Conference on Social, economic, and academic leadership (ICSEAL-6-2019)*, 432–437. https://doi.org/10.2991/assehr.k.200526.062

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based models and influences of time pressure. *Frontiers in behavioral neuroscience*, *11*, 122.

Takaguchi, K., Kappes, A., Yearsley, J. M., Sawai, T., Wilkinson, D. J., & Savulescu, J. (2022). Personal ethical settings for driverless cars and the utility paradox: An ethical analysis of public attitudes in uk and japan. *Plos one*, *17*(11), e0275812.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

UNCTAD. (2023, October). Frontier technology readiness index, annual. https://unctadstat.unctad.org/datacentre/dataviewer/US.FTRI

United Nations Conference on Trade and Development. (2024). Unctadstat - statistical database. https://unctadstat.unctad.org/EN/

United Nations Development Programme. (2024). United nations development programme (undp). https://www.undp.org/

Varoquaux, G., et al. (2023). Joblib: Running python functions as pipeline jobs.

Wang, H., Khajepour, A., Cao, D., & Liu, T. (2020). Ethical decision making in autonomous vehicles: Challenges and research progress. *IEEE Intelligent Transportation Systems Magazine*, *14*(1), 6–17.

Waskom, M. L. (2021). Seaborn: Statistical data visualization.

WHO. (2020, April). Reported distribution of road traffic deaths by type of road user. https://apps.who.int/gho/data/node.main.A998?lang=en

WHO. (2021, February). Road traffic deaths. https://apps.who.int/gho/data/view.main.51310?lang=en

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.

World Bank. (2024). World bank - home page. https://www.worldbank.org/ext/en/home

World Health Organization. (2024). World health organization (who). https://www.who.int/

Yang, C., Brower-Sinning, R. A., Lewis, G., & Kästner, C. (2022). Data leakage in notebooks: Static detection and better processes. *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–12.

Yang, J.-B., Shen, K.-Q., Ong, C.-J., & Li, X.-P. (2009). Feature selection for mlp neural network: The use of random permutation of probabilistic outputs. *IEEE Transactions on Neural Networks*, *20*(12), 1911–1922.

Zhu, A., Yang, S., Chen, Y., & Xing, C. (2022). A moral decision-making study of autonomous vehicles: Expertise predicts a preference for algorithms in dilemmas. *Personality and Individual Differences*, *186*, 111356.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089.

APPENDIX A

Appendix A provides an overview of the libraries that are used in Python (v3.12.7) (Foundation, 2024):

- Pandas (v2.2.3) (McKinney, 2011)
- NumPy (v1.26.4) (Harris et al., 2020)
- PyCountry (v24.6.1) (Ehmann, 2023)
- Scikit-Learn (v1.5.2) (Pedregosa et al., 2011)
- KModes (KPrototypes) (v0.12.2) (Hug, 2017)
- Matplotlib (v3.9.2) (Hunter, 2007)
- Seaborn (v0.13.2) (Waskom, 2021)
- Joblib (v1.4.2) (Varoquaux et al., 2023)
- TensorFlow (v2.16.2) (Abadi et al., 2015)
- Keras (v3.7.0) (Chollet et al., 2015)
- Jinja2 (v3.1.4) (Ronacher, 2023)
- SciKeras (v0.13.0) (DATA-PERSON, n.d.)

APPENDIX B

Appendix B explains all features in the "complete" and "utilitarian" dataset:

- **PedPed:** Every scenario has either pedestrians vs. pedestrians or pedestrians vs. passengers (or passengers vs. pedestrians). This column is 1 if it is pedestrians vs. pedestrians and 0 if it is pedestrians vs. passengers or vice versa (Awad, 2021).
- **Barrier:** This feature indicates whether the potential casualties in this outcome are passengers or pedestrians. It is 1 for passengers and 0 for pedestrians (Awad, 2021).
- **NumberOfCharacters:** This feature represents the total number of characters in this outcome and takes a value between 1 and 5 (Awad, 2021).
- **DiffNumberOfCharacters:** This feature represents the difference in number of characters between this outcome and the other outcome and takes a value between 0 and 4 (Awad, 2021).

- **Man, Woman, Pregnant, Stroller, OldMan, OldWoman, Boy, Girl, Homeless, LargeMan, LargeWoman, Criminal, MaleExecutive, FemaleExecutive, MaleAthlete, FemaleAthlete, MaleDoctor, FemaleDoctor, Dog, Cat:** These features represent the number of character of each type in this outcome. Each feature takes a value between 0 and 5. The sum of these features in one row is always between 1 and 5, the total number of characters in this outcome (Awad, 2021).

- **ICT:** The 'ICT' index indicates the level of ICT infrastructure (UNCTAD, 2023.

- **Skills:** The 'Skills' index indicates the expected years of schooling according to the United Nations Development program and the percentage of working population according to the International Labour Organization (UNCTAD, 2023; United Nations Development Programme, 2024; International Labour Organization, 2024).

- **Research_and_development:** The 'Research and Development' index indicates the number of scientific publications on frontier technologies according to SCOPUS and the number of patents filed on frontier technologies according to PatSeer (UNCTAD, 2023, Scopus, 2024, PatSeer, 2024).

- **Industry_activity:** The 'Industry activity' index indicates the number of products exported and imported with a high R&D intensity as a percentage of the total number of goods imported and exported. It also indicates the number of exports of digitally deliverable services as a percentage of the total services export. The data on these indicators are from UNCTADStat (UNCTAD, 2023, United Nations Conference on Trade and Development, 2024).

- **Finance_access:** The 'Access to finance' index indicates the financial resources provided to the private sector by financial corporations as a percentage of the total GDP according to the World Bank, the International Monetary Fund and the Organisation for Economic Co-operation and Development (UNCTAD, 2023, World Bank, 2024, International Monetary Fund, 2024, Organisation for Economic Co-operation and Development, 2024).

- **Overall_index:** The overall index indicates the country's readiness for using, adopting and adapting frontier technologies (UNCTAD, 2023).

- **Total:** This feature indicates the total estimated number of road traffic deaths per 100,000 population (WHO, 2021).

- **Males:** This feature indicates the estimated number of road traffic deaths for males per 100,000 population (WHO, 2021).
- **Females:** This feature indicates the estimated number of road traffic deaths for females per 100,000 population (WHO, 2021).
- **Passengers:** This feature indicates the estimated number of road traffic deaths for passengers per 100,000 population (WHO, 2020).
- **Pedestrians:** This feature indicates the estimated number of road traffic deaths for pedestrians per 100,000 population (WHO, 2020).
- **ScenarioTypeStrict_Fitness:** This feature indicates that the level of fitness of the characters is different in the two outcomes (Awad, 2021).
- **ScenarioTypeStrict_Gender:** This feature indicates that there are more males in one outcome and more females in the other (Awad, 2021).
- **ScenarioTypeStrict_SocialStatus:** This feature indicates that the social status of the characters is different in the two outcomes (Awad, 2021).
- **ScenarioTypeStrict_Species:** This feature indicates that there are humans in one outcome and pets (cats and/or dogs) in the other (Awad, 2021).
- **ScenarioTypeStrict_Utilitarian:** This feature indicates that the number of characters is different in the two outcomes (Awad, 2021).
- **ScenarioTypeStrict_Random:** This feature indicates that the characters in both outcomes are randomly generated (Awad, 2021).
- **AttributeLevel_Female:** This feature indicates that more characters in this outcome are females compared to the other outcome (Awad, 2021).
- **AttributeLevel_Male:** This feature indicates that more characters in this outcome are males compared to the other outcome (Awad, 2021).
- **AttributeLevel_Fit:** This feature indicates that the characters in this outcome are more fit compared to the other outcome (Awad, 2021).
- **AttributeLevel_High:** This feature indicates that the characters in this outcome have higher social status compared to the other outcome (Awad, 2021).

- **AttributeLevel_Low:** This feature indicates that the characters in this outcome have lower social status compared to the other outcomes (Awad, 2021).
- **AttributeLevel_Hoomans:** This feature indicates that the characters in this outcome are humans (Awad, 2021).
- **AttributeLevel_Pets:** This feature indicates that the characters in this outcome are pets (Awad, 2021).
- **AttributeLevel_More:** This feature indicates that there are more characters in this outcome compared to the other outcome (Awad, 2021).
- **AttributeLevel_Less:** This feature indicates that there are less characters in this outcome compared to the other outcome (Awad, 2021).
- **AttributeLevel_Old:** This feature indicates that the characters in this outcome are older compared to the other outcome (Awad, 2021).
- **AttributeLevel_Young:** This feature indicates that the characters in this outcome are younger compared to the other outcome (Awad, 2021).
- **AttributeLevel_Rand:** This feature indicates that the characters in both outcomes are randomly generated (Awad, 2021).
- **CrossingSignal_1:** This feature indicates that the pedestrians are legally crossing a green traffic light (Awad, 2021).
- **CrossingSignal_2:** This feature indicates that the pedestrians are illegally crossing a red traffic light (Awad, 2021).
- **NumberOfLivesSaved:** This feature indicates the number of net lives that are saved when choosing this outcome. It ranges from minus four (saving one character instead of five) to four (saving five characters instead of one).
- **Finance_nols, ICT_nols, Industry_nols, Overall_nols, Research_nols, Skills_nols, Total_nols, Males_nols, Females_nols, Passengers_nols, Pedestrians_nols:** These feature indicate the interaction terms between the "NumberOfLivesSaved" feature and the features serving as proxy for the availability bias.
- **Clusters_1, Clusters_2, Clusters_3, Clusters_4, Clusters_5, Clusters_6:** These features indicates the number of the cluster this outcome belongs to according to K-prototype Clustering.

APPENDIX C

Model	Hyper-parameters	Best model values
Logistic regression	Solver	liblinear
	Regularization (penalty)	L1
	Regularization strength (C)	0.01
	Maximum iterations	500
	L1 ratio	-
	Tolerance	0.0001
Decision tree	Minimum samples split	20
	Minimum samples leaf	30
	Maximum depth	15
	Criterion	gini
Random forest	Number of trees	250
	Minimum samples split	50
	Minimum samples leaf	20
	Maximum depth	25
	Criterion	entropy
Wiedeman et al. (2020)	Optimizer	rmsprop
	Number of neurons	64
	Number of hidden layers	2
	Activation function	relu
	Number of epochs	50
Neural network	Optimizer	rmsprop
	Number of neurons	16
	Number of hidden layers	3
	Activation function	tanh
	Number of epochs	50

Table 7: Hyper-parameters for the best models for the "complete" dataset.

Model	Hyper-parameters	Best model values
Logistic regression	Solver	lbfgs
	Regularization (penalty)	L2
	Regularization strength (C)	0.01
	Maximum iterations	200
	L1 ratio	-
	Tolerance	0.01
Decision tree	Minimum samples split	30
	Minimum samples leaf	30
	Maximum depth	10
	Criterion	entropy
Random forest	Number of trees	200
	Minimum samples split	30
	Minimum samples leaf	30
	Maximum depth	25
	Criterion	entropy
Wiedeman et al. (2020)	Optimizer	rmsprop
	Number of neurons	64
	Number of hidden layers	2
	Activation function	relu
	Number of epochs	50
Neural network	Optimizer	rmsprop
	Number of neurons	16
	Number of hidden layers	3
	Activation function	tanh
	Number of epochs	50

Table 8: Hyper-parameters for the best models for the "utilitarian" dataset.

APPENDIX D

Figure 15: Confusion matrix for the "utilitarian" test set.

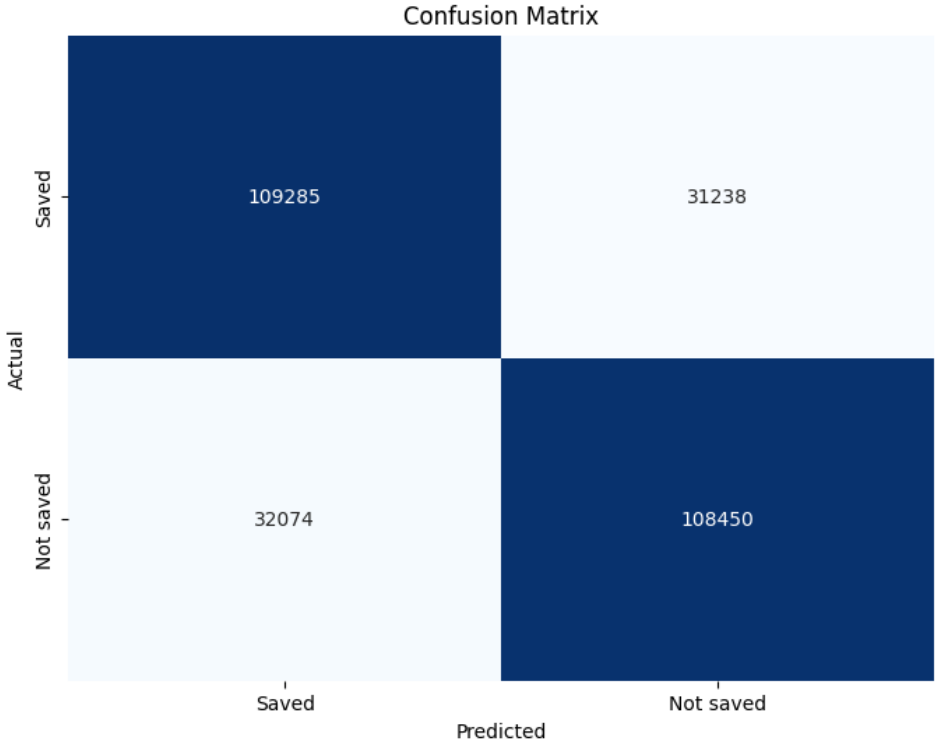
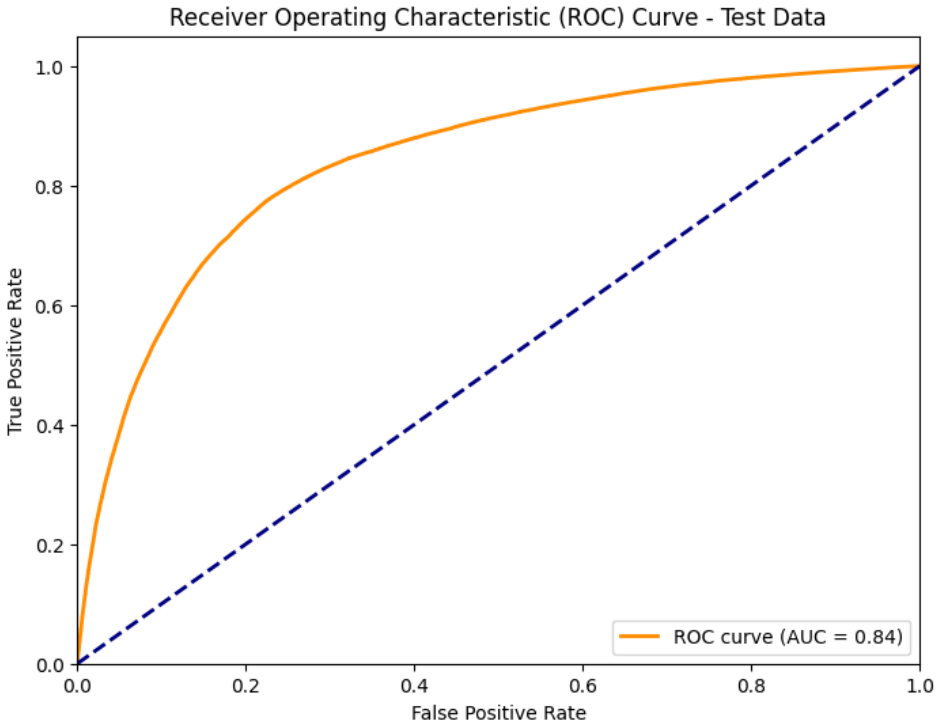


Figure 16: ROC curve for the "utilitarian" test set.



APPENDIX E

Feature	Importance	Relative Importance (%)
Attribute level humans	0.029433	11.389233
Crossing signal 2	0.025978	10.052320
Attribute level more	0.021758	8.419386
Attribute level young	0.020224	7.825967
Crossing signal 1	0.019965	7.725458
Number of characters	0.013848	5.358667
Attribute level pets	0.013731	5.313231
Barrier	0.013528	5.234751
Attribute level high	0.010621	4.109872
Different number of characters	0.010578	4.093350
Attribute level fit	0.009201	3.560512
Attribute level female	0.007785	3.012529
Attribute level old	0.006864	2.655927
Pedestrian pedestrian	0.005885	2.277296
Attribute level less	0.005355	2.072146
Attribute level rand	0.003811	1.474597
Scenario type species	0.003238	1.252926
Total	0.003160	1.222635
Homeless	0.002690	1.040892
Males	0.002014	0.779292
Cat	0.001918	0.742118
Male athlete	0.001498	0.579650
Scenario type gender	0.001416	0.547983
Criminal	0.001288	0.498417
Girl	0.001252	0.484648
Females	0.001245	0.481895
Scenario type fitness	0.001238	0.479141
Scenario type social status	0.001213	0.469503
Old man	0.001196	0.462619
Scenario type utilitarian	0.001131	0.437836
Old woman	0.001043	0.403415
Woman	0.001021	0.395154
Male executive	0.001014	0.392400
Female athlete	0.001011	0.391023
Man	0.000968	0.374501
Female executive	0.000882	0.341457
Boy	0.000836	0.323558
Female doctor	0.000833	0.322181

Feature	Importance	Relative Importance (%)
Dog	0.000765	0.296021
Large man	0.000701	0.271238
Scenario type random	0.000680	0.262977
Overall index	0.000676	0.261600
Pregnant	0.000608	0.235440
Pedestrians	0.000580	0.224425
Level of ICT infrastructure	0.000569	0.220295
Male doctor	0.000569	0.220295
Research and development	0.000455	0.176236
Large woman	0.000423	0.163844
Stroller	0.000331	0.128046
Industry activity	0.000295	0.114278
Clusters 2	0.000260	0.100509
Attribute level low	0.000196	0.075726
Skills	0.000167	0.064712
Attribute level male	0.000149	0.057827
Passengers	0.000149	0.057827
Clusters 1	0.000114	0.044059
Access to finance	0.000068	0.026160

Table 9: Feature importances for all the features in the "complete" dataset.

Feature	Importance	Relative Importance (%)
Skills_nols	0.010827	17.662081
AttributeLevel_More	0.009123	14.881885
CrossingSignal_2	0.007447	12.148122
NumberOfLivesSaved	0.005885	9.600093
CrossingSignal_1	0.005743	9.367926
Barrier	0.003821	6.233676
PedPed	0.002138	3.488305
Total	0.002010	3.279355
Males	0.001968	3.209705
Passengers_nols	0.001427	2.327471
NumberOfCharacters	0.001366	2.228800
Industry_nols	0.001345	2.193975
Total_nols	0.001149	1.874746
Skills	0.001028	1.677404
Overall_nols	0.000904	1.474259
Research_nols	0.000779	1.271113
Males_nols	0.000719	1.172442

Feature	Importance	Relative Importance (%)
DiffNumberOFCharacters	0.000359	0.586221
Finance_nols	0.000352	0.574613
Overall_index	0.000345	0.563004
Stroller	0.000302	0.493354
Man	0.000263	0.429508
MaleDoctor	0.000249	0.406292
Girl	0.000228	0.371467
Clusters_3	0.000221	0.359858
Passengers	0.000196	0.319229
FemaleAthlete	0.000174	0.284404
Clusters_1	0.000153	0.249579
Pregnant	0.000139	0.226363
Boy	0.000135	0.220558
Research_and_development	0.000110	0.179929
Finance_access	0.000103	0.168321
MaleAthlete	0.000096	0.156713
Woman	0.000089	0.145104
LargeMan	0.000089	0.145104
ICT_nols	0.000085	0.139300
Cat	0.000082	0.133496
Pedestrians_nols	0.000075	0.121888
Clusters_5	0.000071	0.116083
Females	0.000071	0.116083
Criminal	0.000060	0.098671
Females_nols	0.000057	0.092867
FemaleExecutive	0.000039	0.063846
LargeWoman	0.000025	0.040629
OldWoman	0.000018	0.029021
FemaleDoctor	0.000014	0.023217
Dog	0.000007	0.011608
Pedestrians	0.000004	0.005804
Homeless	-0.000032	-0.052238
Clusters_2	-0.000046	-0.075454

Table 10: Feature importances for all the features in the "utilitarian" dataset.

APPENDIX F

Figure 17: Learning curve for the best performing neural network on the "complete" dataset.

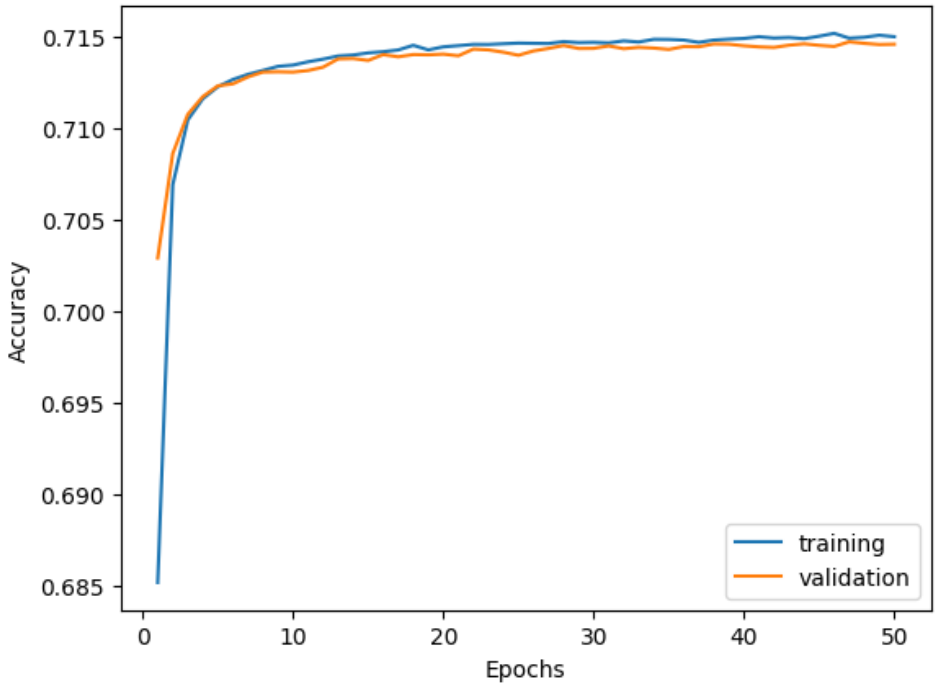
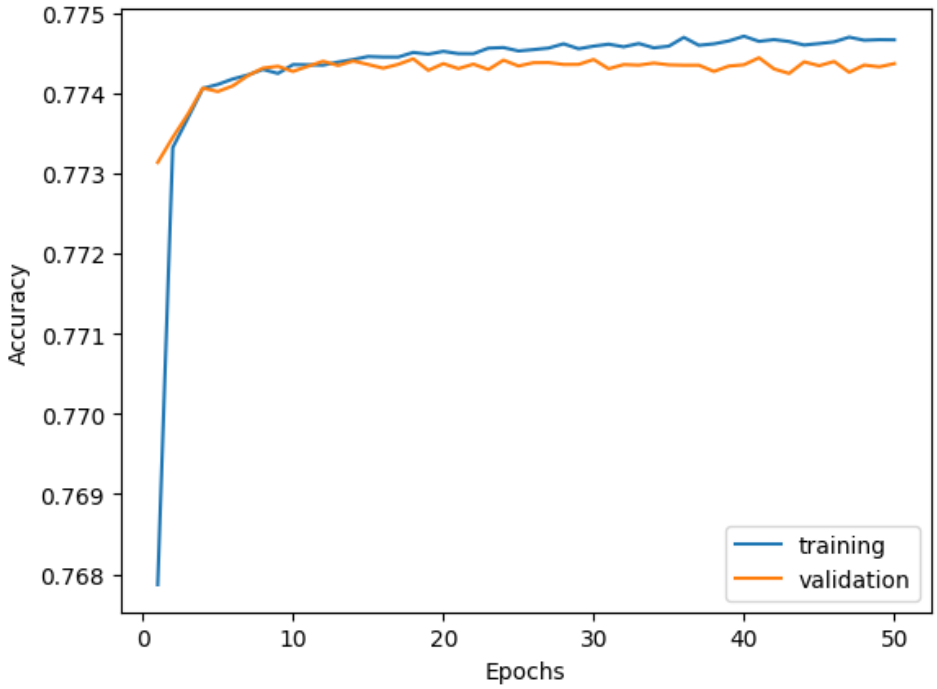


Figure 18: Learning curve for the best performing neural network on the "utilitarian" dataset.



APPENDIX G

Table 11: Descriptive Statistics Table

	count	mean	std	min	25%	50%	75%	max
PedPed	1405234	0.45	0.50	0.00	0.00	0.00	1.00	1.00
Barrier	1405234	0.27	0.45	0.00	0.00	0.00	1.00	1.00
CrossingSignal	1405234	0.61	0.82	0.00	0.00	0.00	1.00	2.00
NumberOfCharacters	1405234	2.91	1.49	1.00	1.00	3.00	4.00	5.00
DiffNumberOFCharacters	1405234	0.51	1.10	0.00	0.00	0.00	0.00	4.00
Saved	1405234	0.50	0.50	0.00	0.00	0.50	1.00	1.00
Man	1405234	0.31	0.59	0.00	0.00	0.00	1.00	5.00
Woman	1405234	0.31	0.59	0.00	0.00	0.00	1.00	5.00
Pregnant	1405234	0.06	0.25	0.00	0.00	0.00	0.00	4.00
Stroller	1405234	0.05	0.24	0.00	0.00	0.00	0.00	4.00
OldMan	1405234	0.17	0.50	0.00	0.00	0.00	0.00	5.00
OldWoman	1405234	0.17	0.50	0.00	0.00	0.00	0.00	5.00
Boy	1405234	0.14	0.43	0.00	0.00	0.00	0.00	5.00
Girl	1405234	0.14	0.43	0.00	0.00	0.00	0.00	5.00
Homeless	1405234	0.13	0.43	0.00	0.00	0.00	0.00	5.00
LargeWoman	1405234	0.15	0.43	0.00	0.00	0.00	0.00	5.00
LargeMan	1405234	0.14	0.43	0.00	0.00	0.00	0.00	5.00
Criminal	1405234	0.05	0.24	0.00	0.00	0.00	0.00	4.00
MaleExecutive	1405234	0.11	0.35	0.00	0.00	0.00	0.00	5.00
FemaleExecutive	1405234	0.11	0.35	0.00	0.00	0.00	0.00	5.00
FemaleAthlete	1405234	0.17	0.50	0.00	0.00	0.00	0.00	5.00
MaleAthlete	1405234	0.17	0.50	0.00	0.00	0.00	0.00	5.00
FemaleDoctor	1405234	0.09	0.32	0.00	0.00	0.00	0.00	5.00
MaleDoctor	1405234	0.09	0.32	0.00	0.00	0.00	0.00	5.00
Dog	1405234	0.16	0.54	0.00	0.00	0.00	0.00	5.00
Cat	1405234	0.16	0.54	0.00	0.00	0.00	0.00	5.00
Finance_access	1405234	0.82	0.10	0.30	0.75	0.85	0.90	1.00
ICT	1405234	0.67	0.13	0.10	0.65	0.65	0.80	1.00
Industry_activity	1405234	0.79	0.08	0.25	0.80	0.80	0.80	1.00
Overall_index	1405234	0.86	0.14	0.10	0.80	0.90	1.00	1.00
Research_and_development	1405234	0.74	0.21	0.00	0.60	0.75	1.00	1.00
Skills	1405234	0.73	0.12	0.00	0.70	0.75	0.75	1.00
Passengers	1405234	3.25	1.59	0.00	1.93	3.46	3.96	19.55
Pedestrians	1405234	1.65	1.10	0.12	0.78	1.44	1.97	16.17
Total	1405234	9.02	5.09	1.75	4.25	7.75	12.50	34.15
Males	1405234	13.65	8.14	2.30	6.65	10.95	17.85	57.90

	count	mean	std	min	25%	50%	75%	max
Females	1405234	8.00	4.78	0.50	4.00	7.00	10.00	33.00

Figure 19: Barplot showing the distribution of the number of times different character types appear across outcomes.

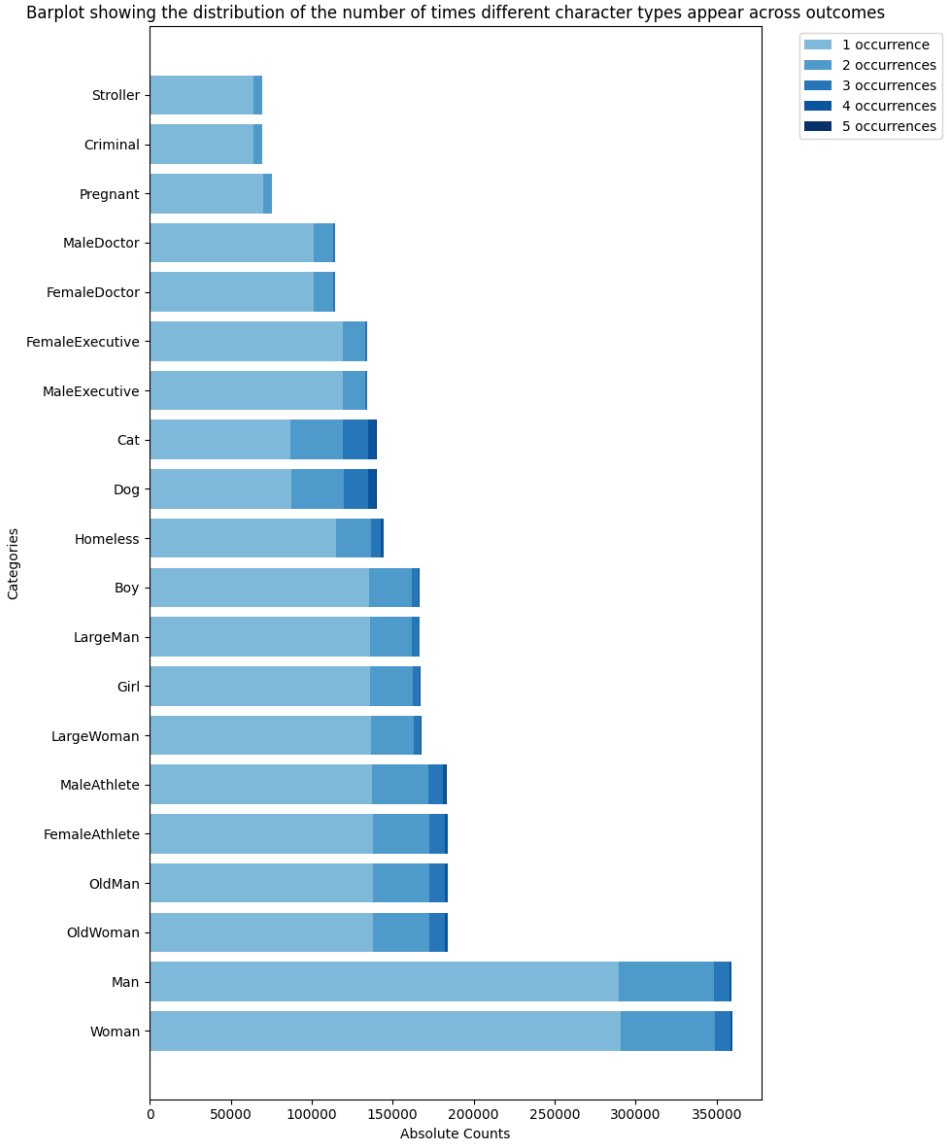


Figure 20: Frequency plot for the statistics on the frontier technology readiness index.

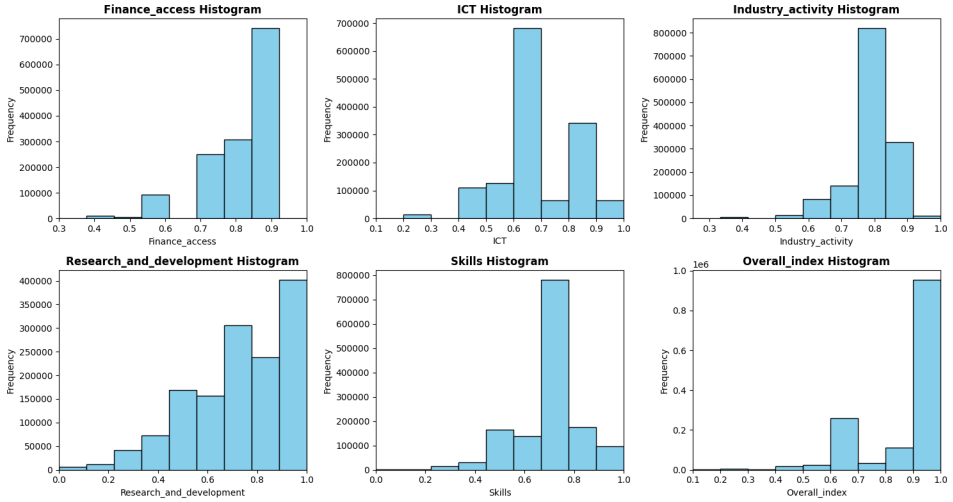


Figure 21: Frequency plot for the road traffic deaths features.

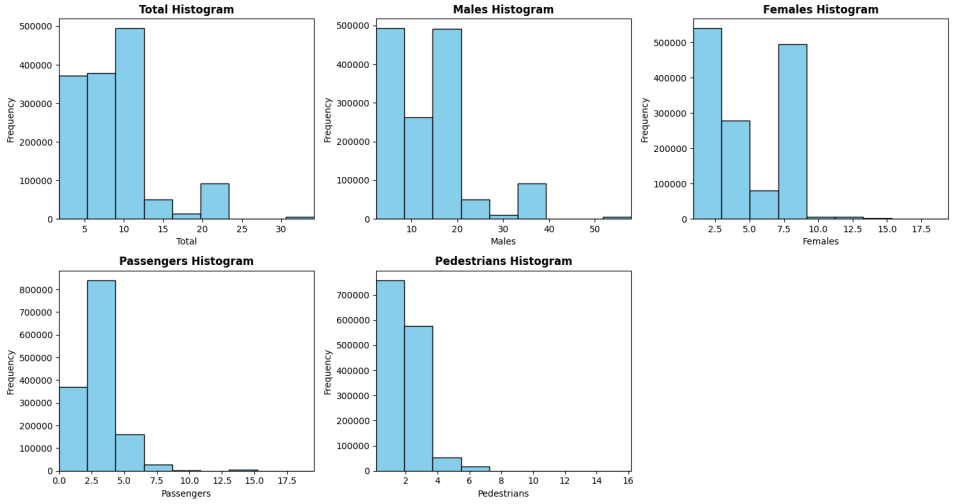


Figure 22: Frequency plot for the different scenario types.

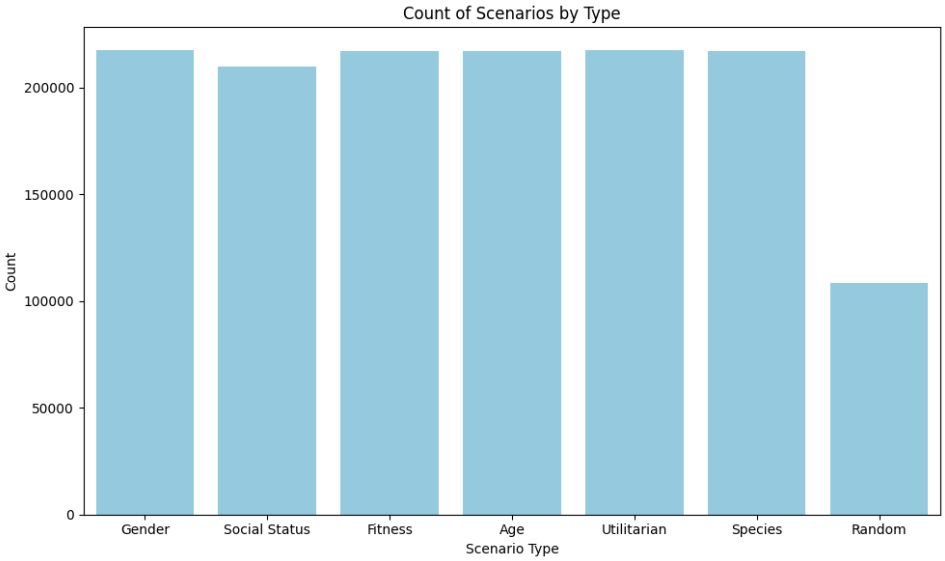


Figure 23: Frequency plot for the different attribute types.

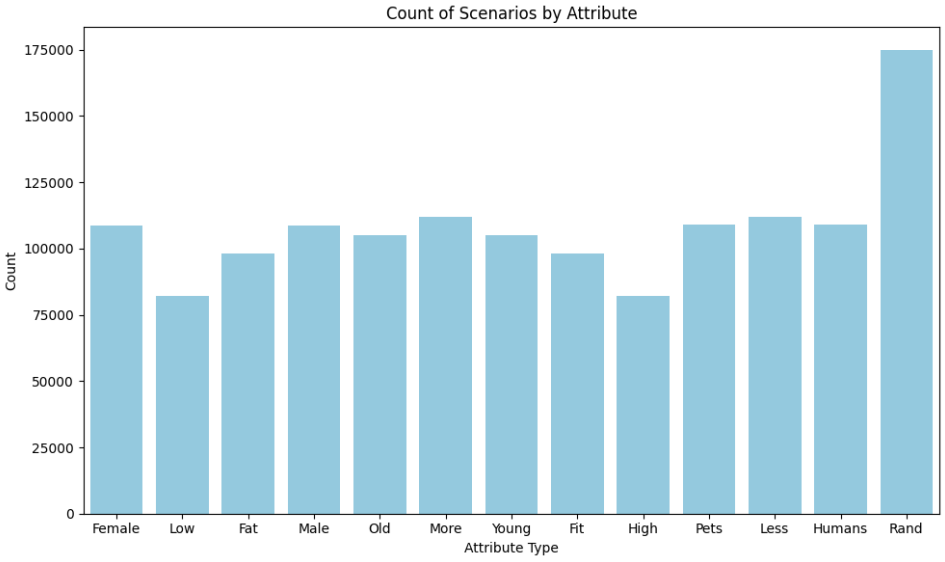


Figure 24: Correlation Matrix showing high correlations between features.

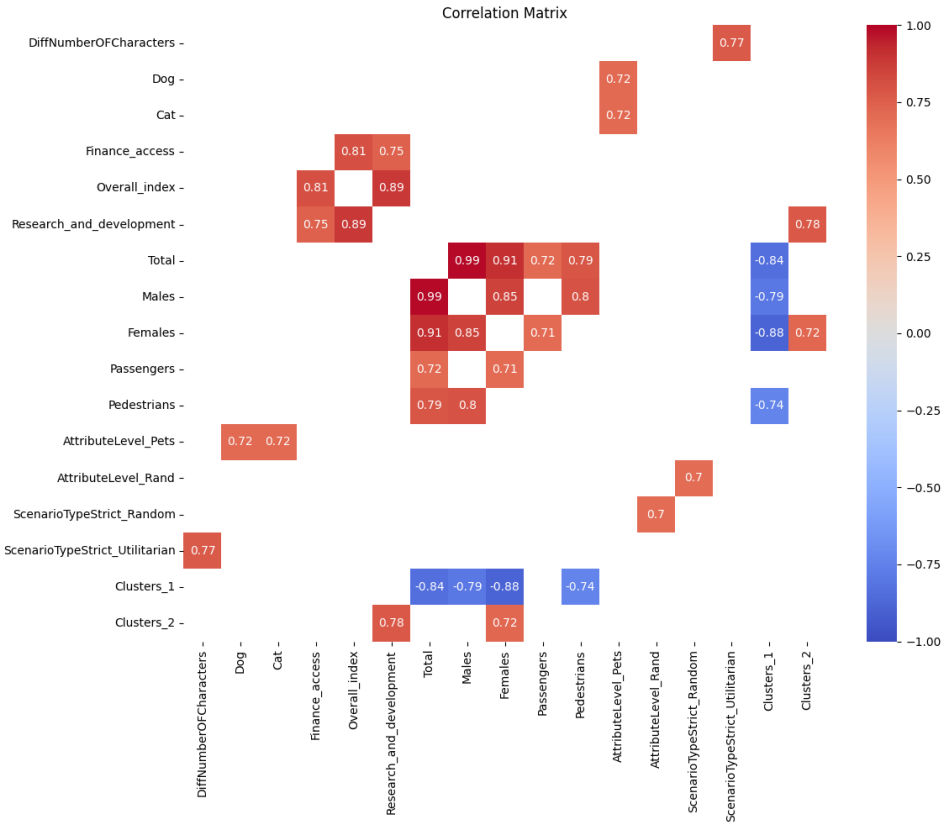


Figure 25: Heatmaps of the interaction effect between the number of lives saved and skills level and between the number of lives saved and the total road traffic deaths.