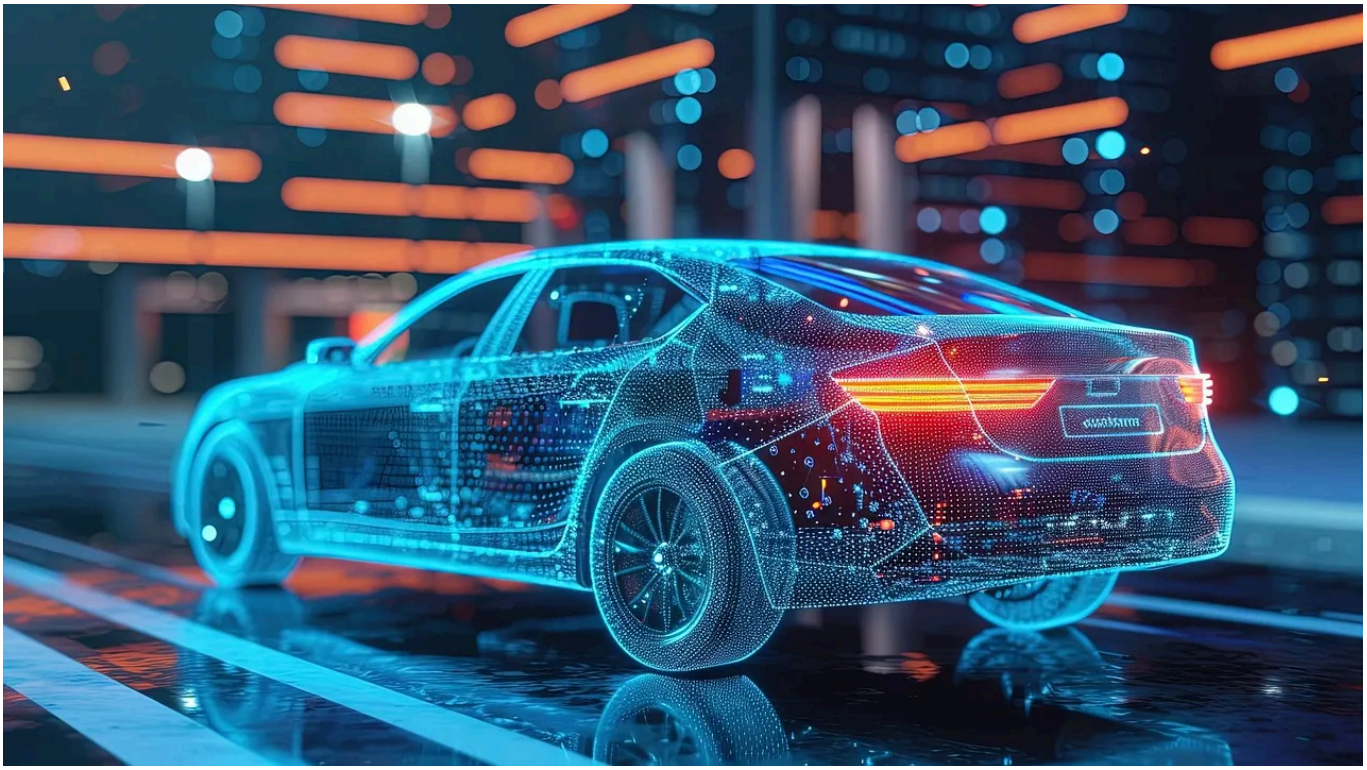
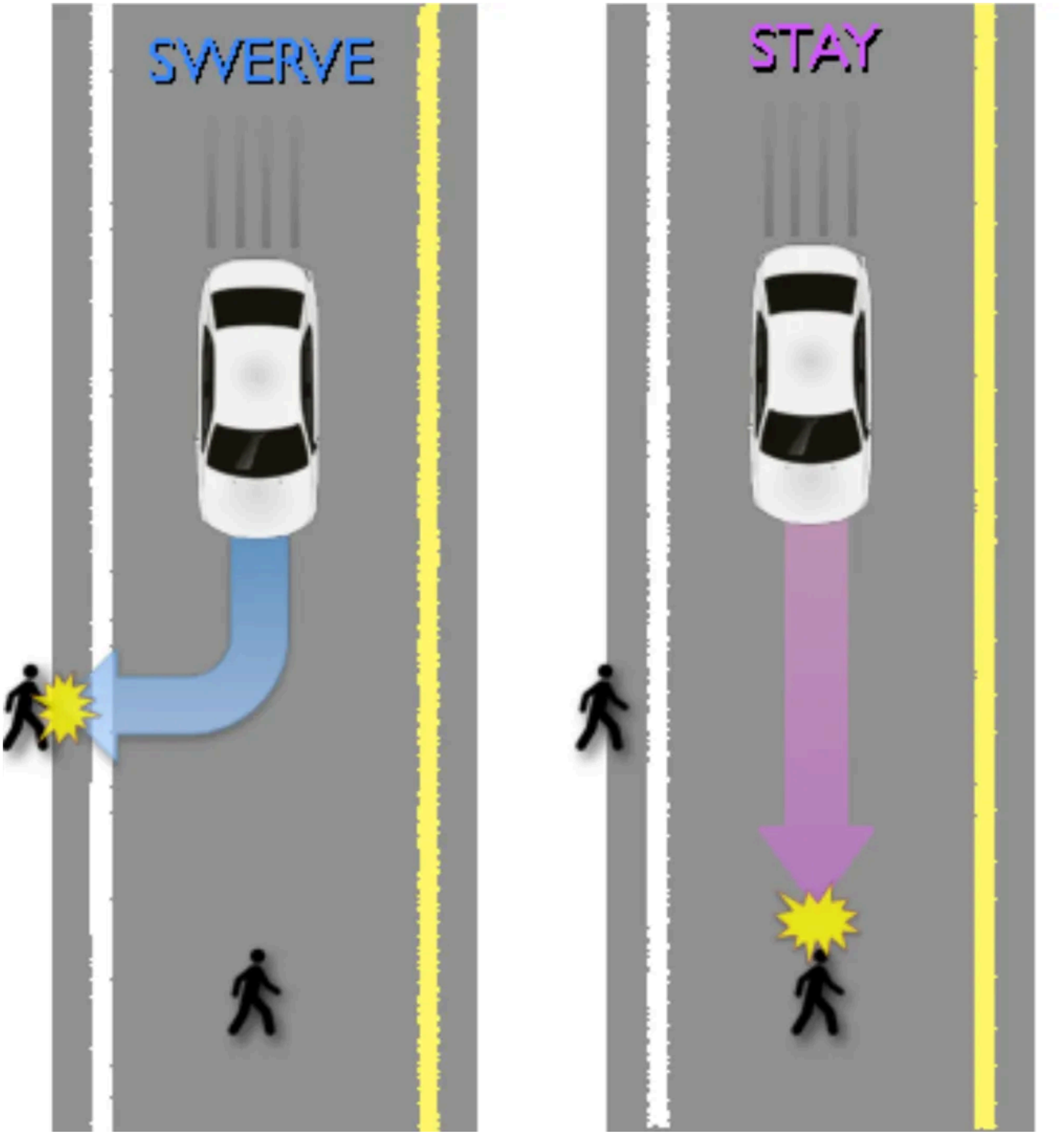


Predicting Human Decision-Making in Autonomous Vehicle Moral Dilemmas

Aileen Li 25 min read · Just now

By: Aileen Li, Alex Edwards, Carol Le, Mikyung Oh, Radha Pawar

Imagine you’re riding in a self-driving car.
The light turns green, the car starts moving, and suddenly a pedestrian jaywalks into the road. To avoid them, the car would have to swerve into two people who are walking legally on the sidewalk. There is no “safe” option, only a choice about *who* is put at risk.



So what would you do? Stay on course, or swerve?

Now imagine the jaywalker was a *little boy*. Would your choice remain the same?

What if the two pedestrians on the sidewalk were a *young man and woman*? Would you still make the same choice?

Millions of people around the world have faced versions of this dilemma in MIT's Moral Machine experiment. In this post, we walk through how we used that data to train machine learning models that predict which outcomes people tend to choose, what patterns emerge in those moral decisions, and what those patterns might mean for real-world AI systems.

Abstract

Autonomous vehicles (AVs) may eventually face situations in which crashes are unavoidable and the system must choose between harmful outcomes. These decisions embed moral trade-offs, such as saving more lives versus obeying traffic laws or prioritizing some groups of people over others. In this project, we use outcome-level data from MIT's Moral Machine experiment to model which AV outcomes humans are most likely to select. Each row represents a possible outcome in a stay–swerve dilemma, along with rich contextual and character features (age, role, legality, species, and more) and a label indicating whether that outcome was chosen.

We compare several supervised learning models, including Logistic Regression, Random Forest, XGBoost, and CatBoost, and evaluate their ability to predict which outcomes are saved. CatBoost achieves the best performance, with ROC–AUC around 0.78 on a held-out test set. We then

examine model behavior across countries, scenario types, and clusters of “moral personas,” and use SHAP values to interpret which features most strongly influence predictions. Overall, we find that human moral decisions are partially predictable and follow consistent patterns (e.g., saving more lives, children over adults, humans over animals, and legal over illegal crossers), but also exhibit substantial variability that no single model fully captures.

1. Introduction & Background

1.1 Problem Motivation

Autonomous vehicles (AVs) will increasingly face situations where crashes are unavoidable and the system must choose between morally significant outcomes, such as saving passengers or pedestrians, prioritizing legal behavior or maximizing lives saved, or deciding between different groups of potential victims. These decisions embed value judgments that humans themselves do not universally agree on.

Understanding how humans make moral decisions in such dilemmas is essential for developing AV systems that align with social expectations, cultural norms, and regulatory standards. However, human preferences are complex, inconsistent, and vary across countries and contexts. This makes AV ethical decision-making a prediction and modeling problem rather than a simple rule-writing problem.

Our project addresses this by modeling patterns in moral decision-making using a large behavioral dataset from the Moral Machine experiment. The goal is to quantify which scenario features (e.g., age, number of lives, legality, species) most influence which outcomes are chosen, and to examine how stable these patterns are across different groups. Although we initially hoped to study how preferences shift over time, limitations in the dataset's temporal information (Section 6.1) prevented robust temporal analysis, so we focus on cross-sectional variation instead.

1.2 Related Work

The Moral Machine experiment (Awad et al., 2018) is the foundational study showing global patterns of ethical preferences in AV scenarios. It highlights cross-cultural clustering in preferences, including tendencies to value the young over the elderly, humans over animals, and larger groups over smaller ones. This work also provides the dataset we analyze in this project.

Ethical AI and AV policy research has explored how public preferences should inform AV decision-making. For example, Bonnefon et al. (2016) describe the "social dilemma of autonomous vehicles," showing that people often want AVs to behave in broadly utilitarian ways but still prefer to ride in cars that prioritize their own safety. This tension illustrates why understanding descriptive moral preferences is important for both policy and product design.

In parallel, interpretability research such as SHAP (Lundberg & Lee, 2017) provides tools to attribute importance to features in complex models. These methods support transparent, post-hoc explanations for predictions, which is crucial in high-stakes ethical domains. Concept drift research (Gama et al., 2014) studies how data distributions and behaviors change over time and provides methods for tracking performance degradation and adapting models to evolving patterns. While our dataset does not support temporal drift analysis, this literature motivated our initial attempt to explore temporal variation.

Together, these strands of work establish the importance of predicting moral decisions, highlight the ethical tensions around AV design, and provide methodological support for our modeling and interpretability approach.

1.3 Approach Overview

Our approach has three main components:

1. Outcome-Level Predictive Modeling
2. Segmentation and Cross-Sectional Analysis
3. Model Interpretability

Outcome-Level Predictive Modeling
We frame the task as a binary classification problem at the outcome level. Each row in the dataset represents one possible outcome in a stay–swerve dilemma and is labeled with `Saved = 1` if that outcome was chosen by the participant, and `Saved = 0` otherwise. We train models to predict the probability that a given outcome will be selected based on its structural and character features. We evaluate a range of models, such as the following:

- Logistic Regression
- Random Forest
- XGBoost
- CatBoost

This combination allows us to compare linear and non-linear models, as well as methods with different handling of categorical features and missing data.

Segmentation and Cross-Sectional Analysis

To understand where a single global model works well and where it breaks down, we evaluate the best model (CatBoost) across different segments. We examine performance by:

- Country or regional group (using `UserCountry3`)
- Scenario type (e.g., Utilitarian, Age, Species, Gender, Fitness, Social Status, Random)
- Unsupervised clusters of outcome types ("moral personas") based on structural and character features

This lets us probe the stability and variability of moral preferences across populations and scenario families without relying on temporal information.

Model Interpretability

Using SHAP values and model coefficients, we quantify how specific attributes influence predictions. We study trade-offs such as:

- Preference for saving more lives
- Preference for lawful versus unlawful pedestrians
- Valuation of human versus non-human entities
- Age- and role-based preferences (e.g., children, elderly, doctors, executives)

Interpretability helps connect model behavior back to ethical considerations and public expectations.

1.4 Contributions and Novel Characteristics

This project makes several contributions:

1. **Outcome-Level Predictive Modeling Framework**
   We develop and evaluate a predictive modeling framework for the Moral Machine dataset at the outcome level, moving beyond descriptive statistics and demonstrating that human moral choices can be learned and generalized to unseen outcomes. This provides a quantitative basis for assessing how predictable moral judgments are in AV-style dilemmas.
2. **Segmentation and Moral Personas**
   We analyze the stability of a global model across countries, scenario types, and unsupervised clusters of outcomes. This cross-sectional analysis reveals where moral preferences appear relatively consistent and where they are more uncertain or context-dependent. Our cluster-based "moral personas" summarize recurring outcome profiles (e.g., children and families, high-status professionals, pets, stigmatized groups) that humans prioritize or deprioritize at different rates.
3. **Interpretable Moral Preference Patterns**
   Using SHAP-based interpretability, we quantify the relative influence of attributes such as age, legality, species, and social status on the probability that an outcome is chosen. This produces transparent, model-driven explanations that connect machine-learning predictions to moral psychology findings and policy-relevant questions.
4. **Business and Societal Implications**
   We discuss how predictive moral modeling can inform the design of AV decision policies, as well as broader applications in marketing, risk assessment, and AI system design. Our results highlight both the potential and the limits of embedding learned moral patterns into real-world systems.

Together, these contributions provide an interpretable, data-driven framework for understanding human moral judgments and their implications for ethically aligned AI systems.

2. Data Collection & Description

2.1 Sources

Our data originates from the Moral Machine experiment (Awad et al., 2018). It is an interactive online platform created by MIT Media Lab to study human moral decisions in AV dilemmas. The full dataset is publicly available through the Open Science Framework under the project repository "Moral Machine Dataset" at https://osf.io/3hvt2/.

The Moral Machine platform presented users worldwide with pairs of scenarios in which an autonomous vehicle must choose between two harmful outcomes: "stay" vs. "swerve." Each scenario describes the number and type of individuals involved, whether they were obeying the laws, and various contextual or structural properties. Users were asked which outcome they preferred. This generated a large-scale behavioral dataset that aggregates judgments from millions of participants across more than 200 countries.

For this project, we focus on **SharedResponse.csv**. This is the primary behavioral log that stores user choices at the outcome level. This file contains approximately 40 million rows, which corresponds to roughly 20 million paired scenarios.

2.2 Methods of Acquisition

Because the full dataset is too large to process within our computational limits, we extracted a **random 100,000-row subset** from the SharedResponse.csv file. This subsample allows us to perform exploratory analysis and model development while keeping computation tractable. Although we did not explicitly enforce stratification or verify representativeness relative to the full dataset, the random sampling procedure preserves broad variability in scenario types, character categories, legality indicators, and user countries. As a result, the subset remains sufficiently diverse for the modeling and interpretability goals of this project, though it may not perfectly reflect the global distribution of the full Moral Machine dataset.

2.3 Dataset Structure

The dataset is organized at the **outcome level**, where each row represents a single possible result of an autonomous vehicle moral dilemma. Every dilemma consists of two outcomes, which correspond to "stay" and "swerve," that share the same `ResponseID`. Together, these two rows reflect one decision made by a participant.

The features describing each outcome include structural and contextual attributes such as `Intervention`, `PedPed`, `Barrier`, `CrossingSignal`, `ScenarioOrder`, `ScenarioTypeStrict`, `ScenarioType`, and `Template`, as well as the participant's country (`UserCountry3`). These fields capture the scenario's setup and the situational constraints under which the moral judgment was made.

Each row also contains detailed character information, listing the count of every character type present in the outcome. The dataset includes more than twenty character categories, such as adults (Man, Woman, Pregnant, LargeMan, LargeWoman), children (Boy, Girl), elderly individuals (OldMan, OldWoman), professionals or roles (MaleExecutive, FemaleExecutive, MaleDoctor, FemaleDoctor, MaleAthlete, FemaleAthlete), vulnerable groups (Homeless, Criminal, Stroller), and animals (Dog, Cat). The variable `NumberOfCharacters` summarizes the total number of individuals affected in that outcome. These character features represent morally relevant distinctions that participants may consider when choosing between outcomes.

The column `Saved` indicates whether the characters in that specific outcome were spared according to the participant's decision. Exactly one of the two rows per `ResponseID` has `Saved = 1`. This structure reflects a **relative choice**: the label describes which of the two alternatives was preferred, not an evaluation of the row in isolation. In our modeling, we use these outcome-level rows directly and treat `Saved` as the target label, allowing the classifier to learn which types of outcomes are more likely to be chosen based on their attributes.

3. Data Pre-Processing & Exploration

3.1 Data Preprocessing

Before fitting predictive models, we performed several preprocessing steps on the outcome-level dataset. Although each dilemma in the Moral Machine platform is presented as a pair of possible outcomes ("stay" and "swerve"), we intentionally chose to model the data at the **outcome level**, using each row as an independent observation. This approach treats the target variable `Saved` as an indicator of whether a given outcome was selected by the participant. Modeling at the outcome level allows the classifier to learn which attributes are associated with outcomes that tend to be chosen, without reconstructing paired records or computing difference features.

This design choice preserves the original structure of the data and simplifies the modeling pipeline while still supporting interpretable feature analyses.

To ensure fair evaluation and preserve the class distribution of the target variable, we split the dataset into stratified training and testing sets as shown below.

```
df = pd.read_csv("ethicaldilemma.csv")

target = "Saved"
features = [
    "PedPed", "Barrier", "CrossingSignal", "ScenarioTypeStrict",
    "NumberOfCharacters", "Template", "DescriptionShown", "LeftHand",
    "Man", "Woman", "Pregnant", "Stroller", "OldMan", "OldWoman",
    "Boy", "Girl", "Homeless", "LargeWoman", "LargeMan", "Criminal",
    "MaleExecutive", "FemaleExecutive", "FemaleAthlete", "MaleAthlete",
    "FemaleDoctor", "MaleDoctor", "Dog", "Cat"
]

X = df[features]
y = df[target].astype(int)

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)
```

The dataset contains substantial missingness across scenario attributes, reflecting the diversity of scenario templates rather than data corruption. Because missingness often carries meaningful information about scenario structure, we retained missing values rather than applying imputation. Tree-based models such as CatBoost can handle missing values natively; for other model families, missingness is represented explicitly through one-hot encoded categorical indicators.

To prepare the dataset for models requiring fully numerical input, we applied one-hot encoding to categorical features such as `ScenarioTypeStrict`, `Intervention`, and `CrossingSignal`. This ensured consistent representation across models. In addition, we standardized numerical features for models sensitive to feature scaling, including logistic regression and neural networks, to improve optimization stability. Tree-based models such as Random Forest and XGBoost were trained on unstandardized data, as these algorithms are invariant to monotonic transformations of feature scales.

These preprocessing steps resulted in a clean, model-ready outcome-level dataset that supports our goal of identifying which scenario attributes and character features are associated with outcomes that humans are more likely to select. The prepared data serves as the foundation for the predictive modeling and interpretability analyses described in the following sections.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the structure of the outcome-level dataset, characterize the diversity of scenario attributes, and examine preliminary moral preference patterns prior to model fitting. Because the dataset consists of a **random sample** of 100,000 outcome rows from SharedResponse.csv, EDA helps verify that the subset preserves essential variability in scenario types, character distributions, and decision outcomes.

Scenario Type Distribution.
We first examined the distribution of `ScenarioTypeStrict` categories, which capture broad ethical themes in Moral Machine dilemmas such as Utilitarian, Age, Species, Gender, Fitness, Social Status, and Random scenarios. As shown in **Figure 1**, Utilitarian scenarios are the most frequent, followed by Age and Species dilemmas. Social Status and Gender scenarios

appear less often but show higher variance. This imbalance reflects the underlying structure of the Moral Machine platform, which emphasizes classic ethical comparisons such as saving more people or prioritizing younger over older individuals.

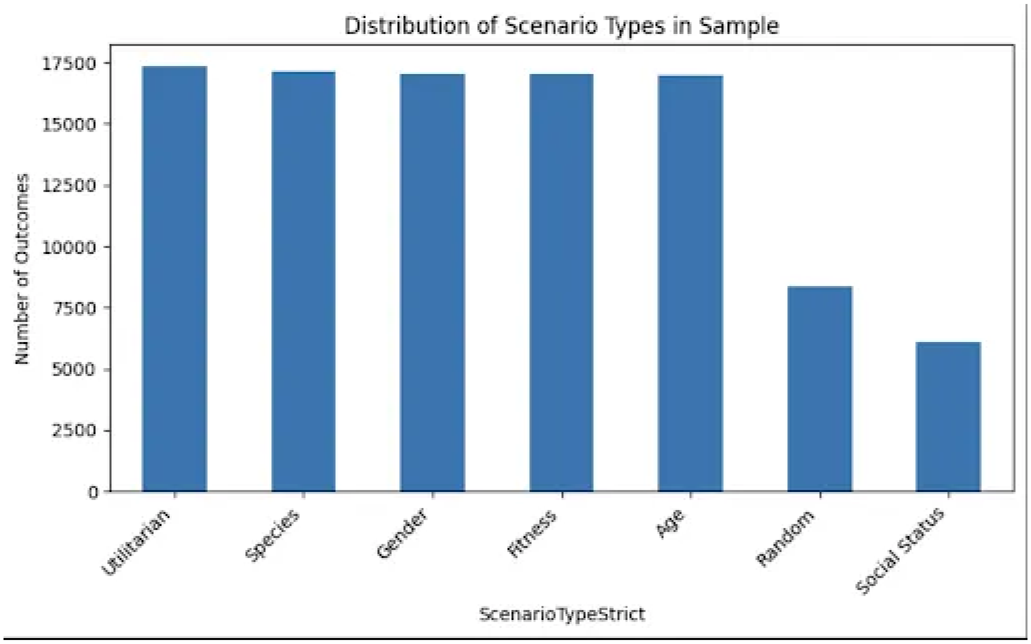


Figure 1. Distribution of high-level scenario types (`ScenarioTypeStrict`) in the 100,000-row outcome sample. Utilitarian cases dominate, followed by Age and Species dilemmas.

Character Frequency Across Outcomes.

Character-count features reveal how often each morally relevant group appears across outcomes. We aggregated raw character categories into interpretable groups (children, adults, elderly, professionals, athletes, low-social-status groups, and animals). **Figure 2** shows that adults and children are the most common groups, while elderly individuals, pets, and status-related groups appear less frequently. These distributions are important because they influence both model training and SHAP-based interpretability: more frequently appearing groups exert greater influence on learned decision boundaries.

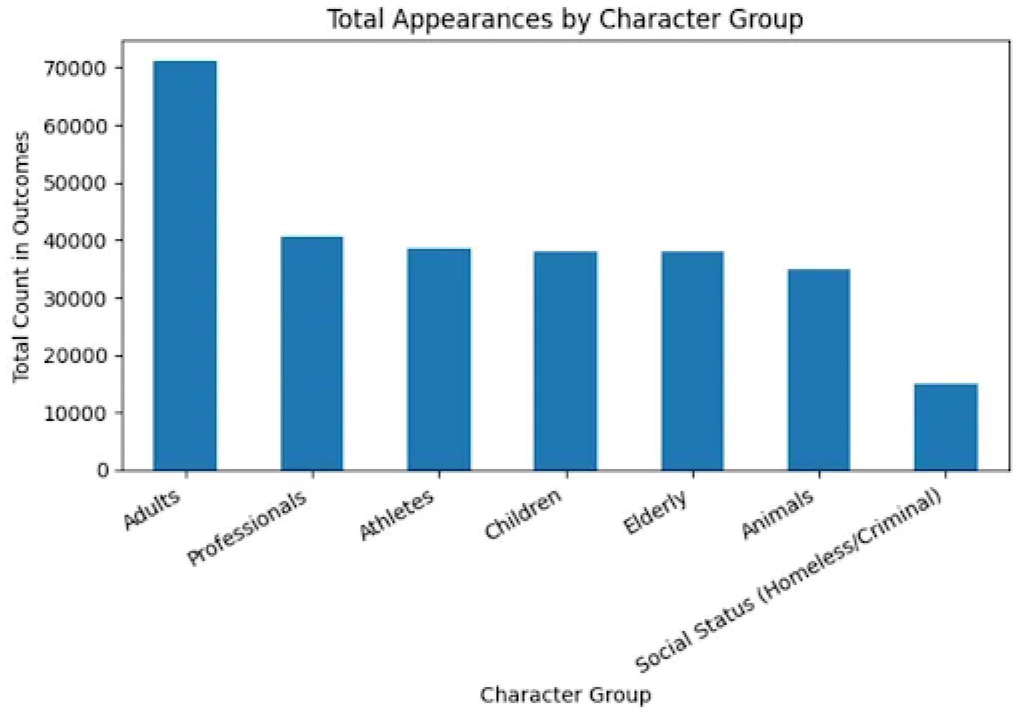


Figure 2. Frequency of character groups across all outcomes. Adults and children appear most often, while elderly individuals, pets, and social-status groups appear less frequently.

Saved vs. Not-Saved Distribution.

We examined the baseline distribution of the target variable `Saved`. **Figure 3** shows a mild imbalance, with the positive class (`Saved = 1`) occurring slightly more often. This reflects a weak but consistent preference for one side of the dilemma across scenarios. Since the imbalance is relatively small, standard classifiers remain stable without requiring explicit resampling.

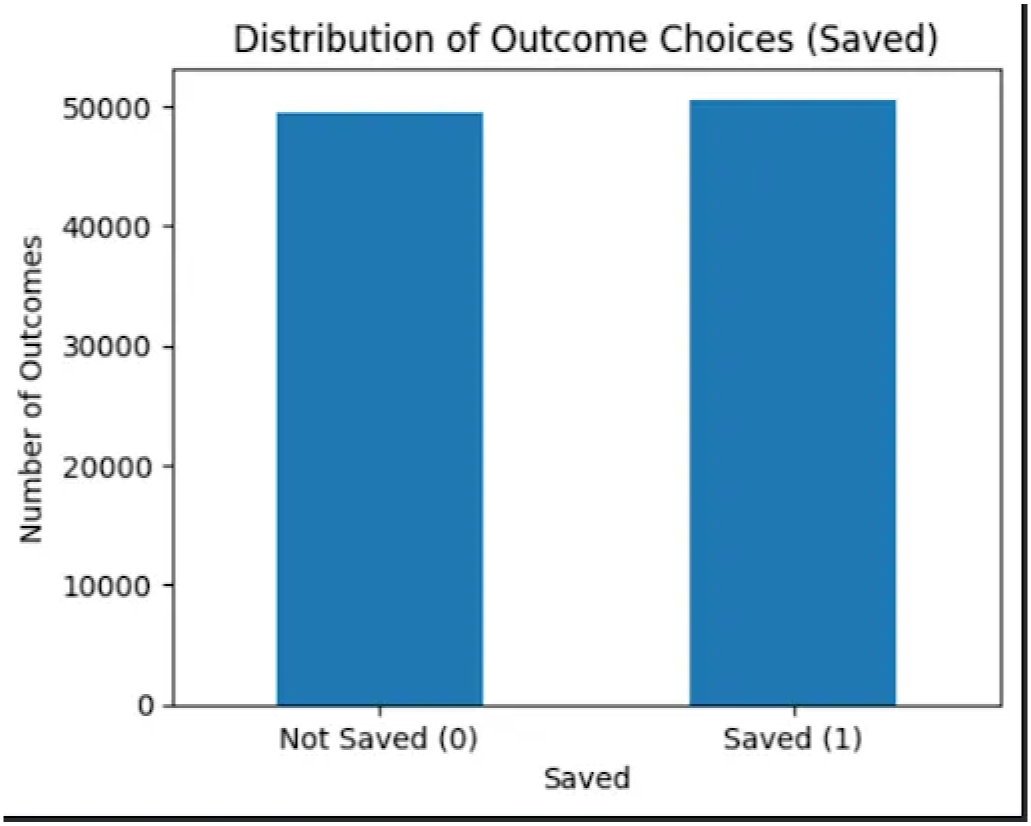


Figure 3. Baseline distribution of the target variable `Saved`. A slight imbalance indicates marginally more outcomes being selected across scenarios.

Legality and Moral Preference Patterns.

Legality is one of the strongest predictors of moral choice in Moral Machine experiments. To investigate this, we computed the empirical probability that an outcome is chosen (`Saved = 1`) conditional on `CrossingSignal`. As shown in **Figure 4**, outcomes involving legally crossing pedestrians have substantially higher saved rates, whereas jaywalking scenarios reduce the likelihood of being chosen. This aligns with prior literature indicating a strong norm-compliance bias in human moral reasoning.

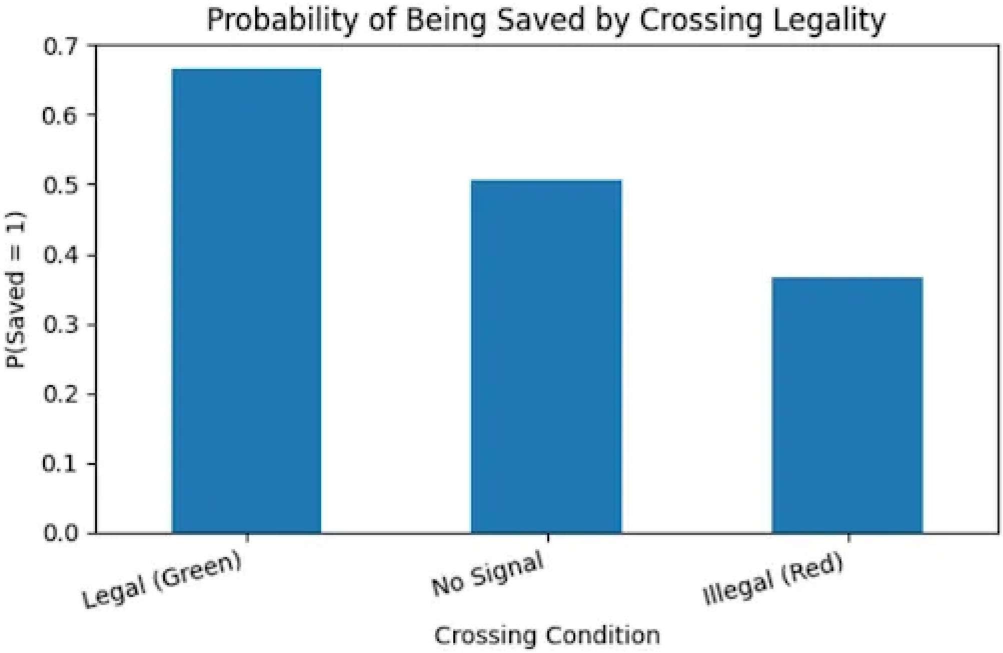


Figure 4. Saved rate conditional on crossing legality (`CrossingSignal`). Legally crossing pedestrians are chosen at substantially higher rates relative to jaywalkers.

Early Moral Patterns in the Raw Data.

Even before applying machine learning models, the EDA reveals several stable ethical tendencies:

- A preference for saving the larger group rather than the smaller group (utilitarian preference).
- Strong prioritization of children and younger individuals relative to adults and the elderly.
- A consistent preference for saving humans over animals.
- Elevated saved rates for high-social-value groups (doctors, executives, athletes).
- Lower saved rates for socially stigmatized groups (criminals, homeless individuals).

These early patterns preview many of the SHAP findings observed later in the Results section, suggesting that the dataset encodes predictable, structured moral preferences even before fitting any predictive model.

4. Modeling Approach

4.1 Model Selection and Rationale

To model the probability that a given **outcome** is chosen (`Saved = 1`), we evaluated four supervised learning algorithms: Logistic Regression, Random Forest, XGBoost, and CatBoost. These models were chosen to balance interpretability, predictive strength, and the ability to handle heterogeneous feature types.

- **Logistic Regression** served as a baseline model due to its simplicity and interpretability. It provides a clear reference point to benchmark non-linear models.
- **Random Forest** was selected for its ability to model non-linear interactions and its robustness to noise and feature scaling.
- **XGBoost** was included because it typically offers strong performance on tabular data and handles complex interactions efficiently through gradient-boosted decision trees.
- **CatBoost** was chosen because it natively handles categorical variables without requiring one-hot encoding, manages missing data effectively, and tends to perform well with minimal hyperparameter tuning.

These models represent a diverse set of learning paradigms, allowing us to compare how linear, bagged-tree, and boosted-tree approaches handle moral judgment prediction.

4.2 Segmentation Design: Stability of the Global Model

In addition to comparing different classifiers, we also want to understand where a single global model works well and where it begins to break down. To do this, we treat segmentation as part of our modeling approach. We first train a global CatBoost model on all users and scenarios (Section 4.1), and then probe its behavior across different populations and types of dilemmas.

Methodologically, we keep the trained CatBoost model fixed and re-evaluate it on different subsets of the held-out test set. This lets us ask:

- Does the same model perform similarly across countries with different cultures?
- Are some types of scenarios inherently easier or harder to predict?
- Can we discover latent “moral personas” purely from the structure and cast of the scenario?

We implement two complementary strategies:

1. **Supervised segmentation**, based on known labels such as country and scenario type.
2. **Unsupervised segmentation**, using K-means clustering on structural and character-count features.

The resulting segment-level performance is visualized in bar charts and tables referred to below.

4.2.1 Supervised Segmentation: Country and Scenario Type

For supervised segmentation, we reuse the global CatBoost model from Section 4.1 and partition the test set along two interpretable axes:

- **Country (`UserCountry3`):** three-letter country code of the respondent.
- **Scenario type (`ScenarioTypeStrict`):** high-level category of the dilemma (Utilitarian, Age, Species, Gender, Fitness, Social Status, Random).

We implement a helper function that:

- Groups the test data by a chosen segment column (e.g., country or scenario type).
- Discards groups with fewer than a minimum size (we use 300 test observations) to avoid unstable metrics.
- For each remaining group, constructs the feature matrix using the same outcome-level features as the global model, ensures categorical variables are cast to strings with missing values filled as `"NA"`, and calls `model.predict_proba` to obtain probabilities.
- Computes accuracy, precision, recall, F1, and ROC–AUC within each group.

Formally, for each group g, we evaluate all metrics on the subset {(x_i, y_i): i ∈ D_g} using the same decision rule p_hat(x) learned on the full training data. We do **not** train separate models per group; segmentation is purely a post-hoc stress test of the global model.

We then visualize segment-level ROC–AUC using bar charts:

Country segmentation. After grouping by `UserCountry3`, we sort countries by ROC–AUC and keep the ten largest segments. The resulting ROC–AUC values are plotted in a bar chart (**Figure 5**), with country codes on the x-axis and ROC–AUC on the y-axis.

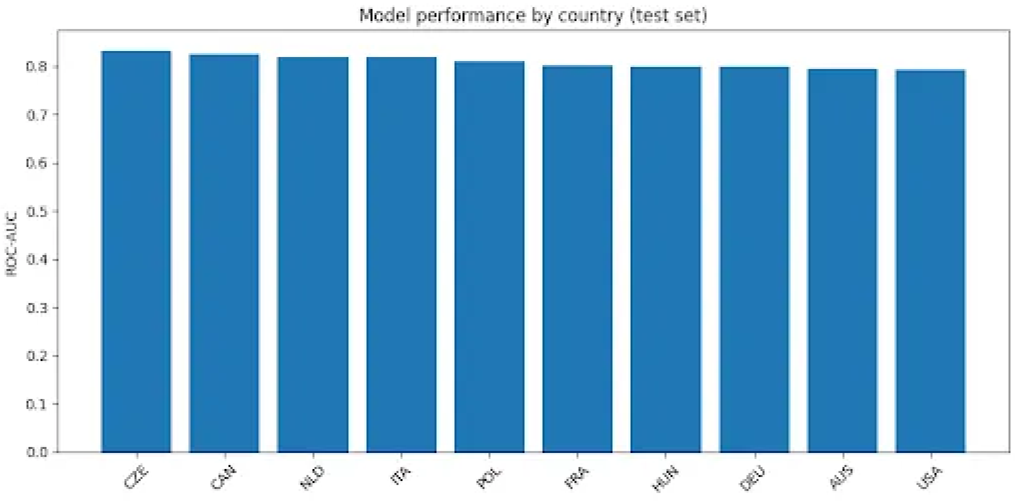


Figure 5. Model performance (ROC–AUC) by country on the test set.

Scenario-type segmentation. We repeat the same procedure using `ScenarioTypeStrict` as the grouping variable. The ROC–AUC for each scenario family (Utilitarian, Age, Species, Gender, Fitness, Social Status, Random) is visualized in a second bar chart (**Figure 6**).

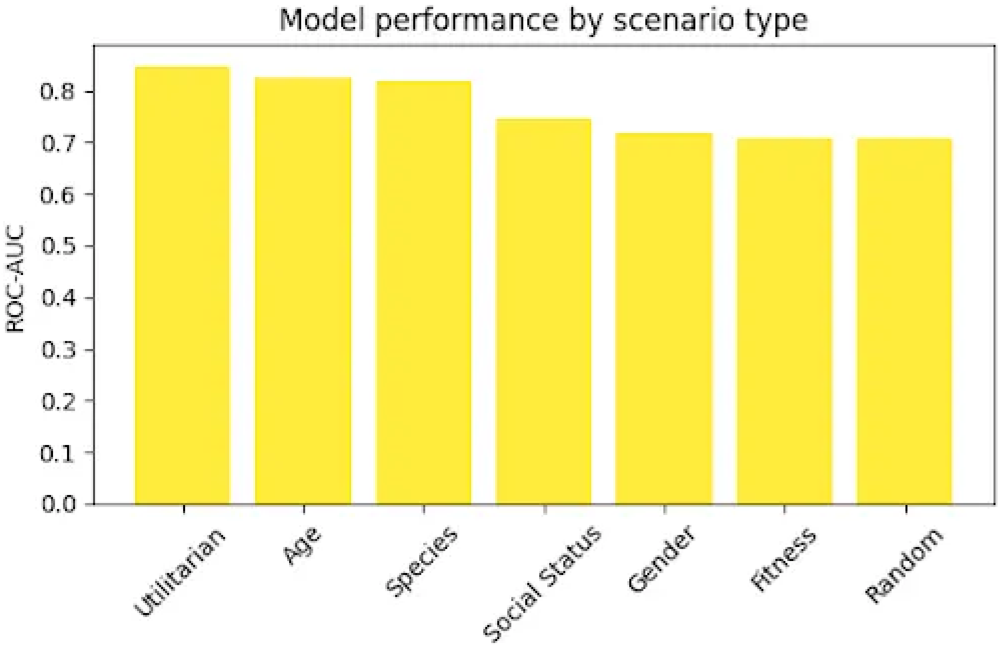


Figure 6. Model performance (ROC–AUC) by scenario type (`ScenarioTypeStrict`) on the test set.

These two figures define the segmentation questions that we later interpret in the Results section: country vs. scenario-type variation in model performance.

4.2.2 Unsupervised Segmentation: K-Means Moral Personas

Supervised segmentation relies on labels we already know. To uncover whether the data itself suggests natural groupings of dilemmas, we also perform unsupervised segmentation using K-means clustering.

```
X_full = preprocessor.transform(X)

kmeans = KMeans(n_clusters=8, random_state=42)
clusters = kmeans.fit_predict(X_full)

df["Cluster"] = clusters
```

Feature construction and preprocessing.
We build a feature matrix that captures both scenario structure and character composition at the outcome level:

- **Structural features:** `PedPed`, `Barrier`, `CrossingSignal`, `ScenarioTypeStrict`, `NumberOfCharacters`, `Template`, `DescriptionShown`, `LeftHand`.
- **Character-count features:** `Man`, `Woman`, `Pregnant`, `Stroller`, `OldMan`, `OldWoman`, `Boy`, `Girl`, `Homeless`, `LargeWoman`, `LargeMan`, `Criminal`, `MaleExecutive`, `FemaleExecutive`, `FemaleAthlete`, `MaleAthlete`, `FemaleDoctor`, `MaleDoctor`, `Dog`, `Cat`.

Categorical variables (`PedPed`, `Barrier`, `CrossingSignal`, `ScenarioTypeStrict`, `Template`, `DescriptionShown`, `LeftHand`) are one-hot encoded, and numeric variables are standardized using a `ColumnTransformer` combined with `StandardScaler`. K-means is then fit on this transformed feature space.

We experiment with k=3, 4, … , 9 and examine both inertia and silhouette score. Inertia decreases monotonically with k, while the silhouette score improves up to around k=8 and only slightly further at k=9. Based on this elbow-plus-silhouette pattern, we set k=8 as a compromise between cluster quality and interpretability (**Figure 7**).

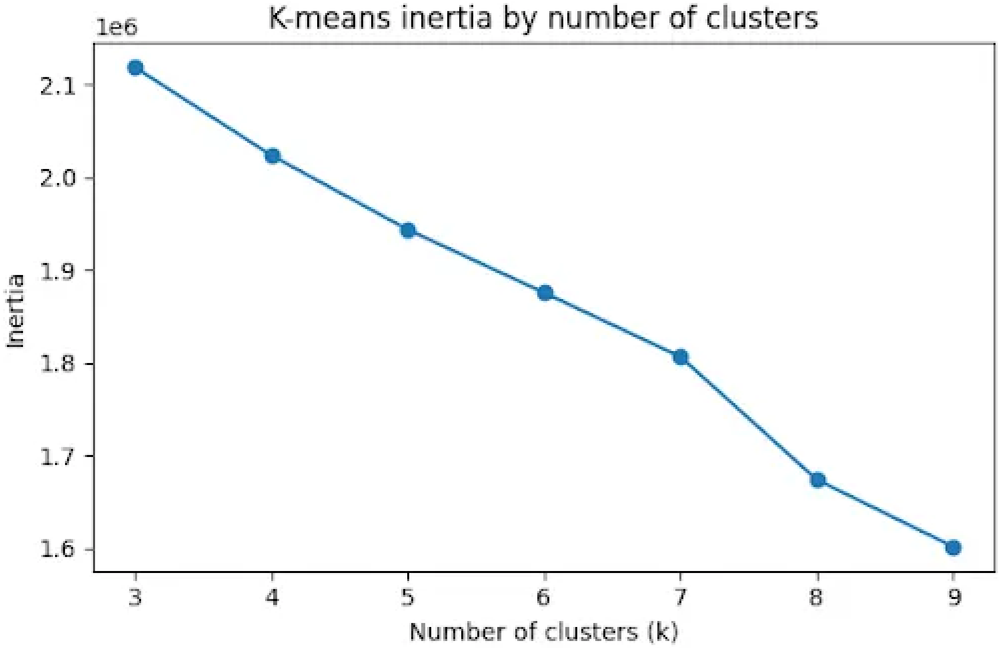


Figure 7. Inertia and silhouette score as a function of the number of clusters kkk. k=8k = 8k=8 is chosen as a balance between fit and interpretability.

Cluster-level summaries.
After assigning each outcome to one of the eight clusters, we summarize the clusters along two key dimensions:

- **Saved_Rate:** the average value of `Saved` within the cluster, which can be interpreted as the empirical probability that outcomes of this type are chosen.
- **Num_Cases:** the total number of outcomes assigned to the cluster, indicating how common that scenario profile is in the data.

For each cluster, we compute these two quantities and report them in a summary table (**Table 1**) and a bar chart (**Figure 8**).

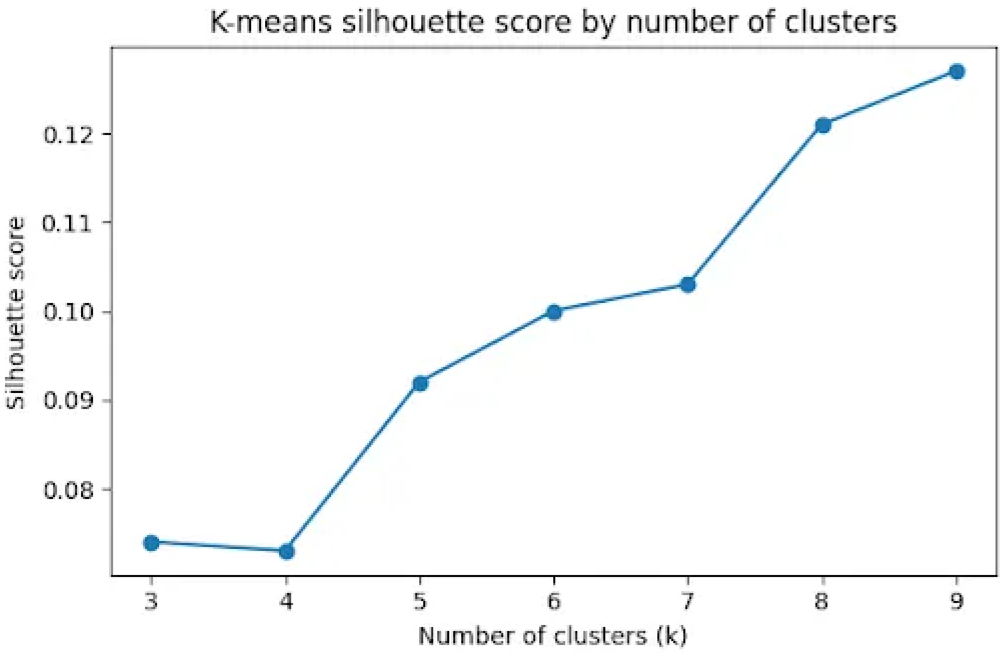
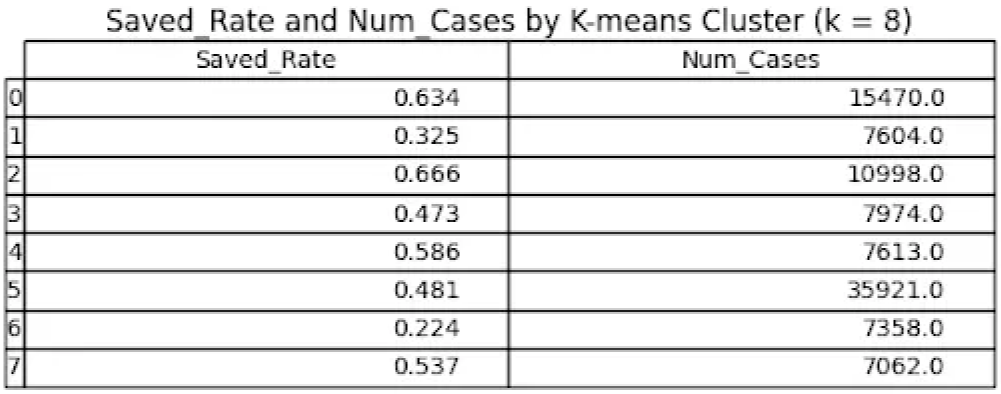


Figure 8. Saved_Rate by K-means cluster (k = 8).



Saved_Rate and Num_Cases by K-means Cluster (k = 8)

	Saved_Rate	Num_Cases
0	0.634	15470.0
1	0.325	7604.0
2	0.666	10998.0
3	0.473	7974.0
4	0.586	7613.0
5	0.481	35921.0
6	0.224	7358.0
7	0.537	7062.0

Table 1. Saved_Rate and Num_Cases by K-means cluster (k = 8).

Interpreting clusters as moral personas.

To understand what each cluster represents, we compute the mean of all structural and character-count features within each cluster and examine the features with the highest average values. Looking at these dominant features reveals intuitive patterns:

- Some clusters are dominated by children and parents.
- Others are rich in doctors and executives.

- Some contain mostly athletes.
- Others are dominated by elderly characters.
- Some clusters consist primarily of pets (dogs and cats) or homeless individuals.

These feature profiles allow us to interpret the clusters as latent "**moral personas**": recurring scenario types that humans tend to save or sacrifice at different rates. In the Results section, we use these personas to interpret Table 1 and Figure 8, showing how different moral personas occupy different positions in the hierarchy of human moral concern in AV dilemmas.

4.3 Training Methods, Validation, and Parameter Selection

A consistent training procedure was used to ensure fair model comparison. The dataset was split into 80% training and 20% testing, stratified on the target variable to preserve the distribution of `Saved` vs. `Not Saved`.

For Logistic Regression, Random Forest, and XGBoost, we used a `sklearn Pipeline` that combined preprocessing with model fitting to maintain clean, reproducible workflows. Categorical variables were transformed using one-hot encoding, and numerical variables were scaled using `StandardScaler`.

Parameter choices were guided by two principles:

- **Stability and reproducibility** rather than heavy optimization
- **Avoiding overfitting**, given the moderate dataset size

For this reason, we used:

- Logistic Regression with increased iteration limits
- Random Forest with default depth and reproducible randomness
- XGBoost with `eval_metric="logloss"` to prioritize proper probability calibration
- CatBoost with 1000 iterations and automatic best-iteration selection on a validation set

CatBoost was trained separately because it handles categorical variables and missing values internally and does not require preprocessing through a `ColumnTransformer`. Below is the core training routine used to fit the CatBoost classifier on the outcome-level dataset.

```
cat_model = CatBoostClassifier(
    iterations=1000,
    loss_function="Logloss",
    eval_metric="AUC",
    random_state=42,
    verbose=False
)

cat_model.fit(
    X_train, y_train,
    cat_features=cat_cols,
    eval_set=(X_test, y_test),
    use_best_model=True
)
```

All models were evaluated on the held-out test set to provide an unbiased estimate of real-world performance.

5. Results

5.1 Key Findings and Evaluation

Across all models, CatBoost achieved the strongest performance, particularly on ROC–AUC (0.776), followed closely by XGBoost (0.773). Logistic Regression underperformed relative to tree-based methods, suggesting that non-linear interactions are critical in predicting which outcomes humans choose.

CatBoost also produced the most stable probability estimates and required the least preprocessing, making it the most efficient and reliable model for this dataset. Precision and recall scores were consistent across models, averaging near **0.71** for the top performers. This indicates that the models are relatively balanced when predicting which outcomes will be saved.

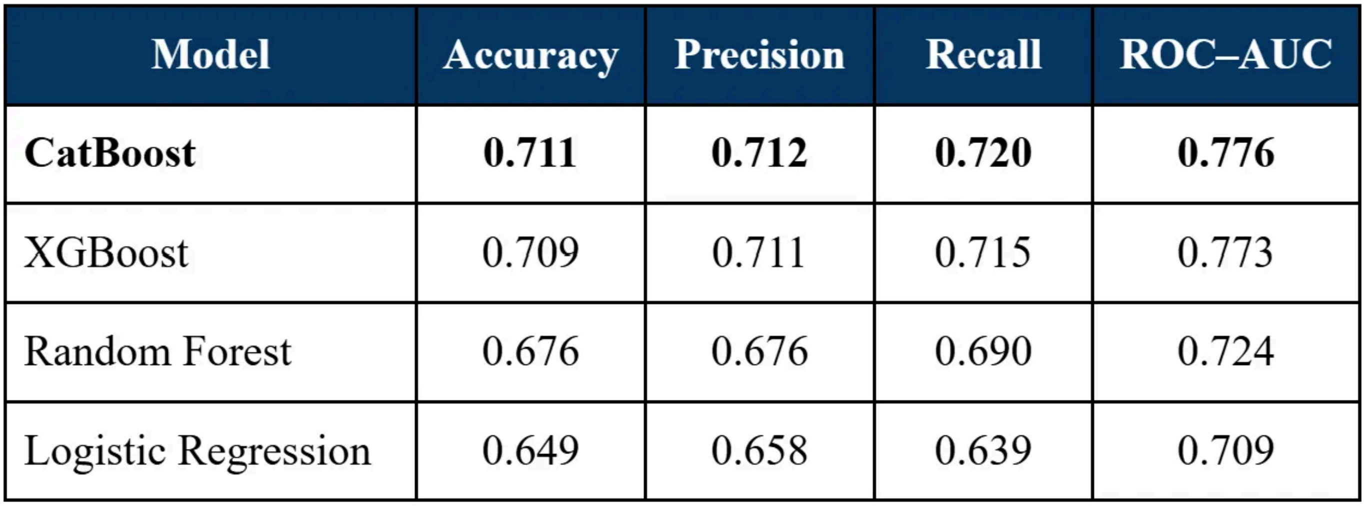
5.2 Comparison Across Models

After training, we evaluate the model on the held-out test set using standard classification metrics.

```
y_pred = cat_model.predict(X_test)
y_prob = cat_model.predict_proba(X_test)[:, 1]

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_prob)
```

A summary of the test results is provided below:



Model	Accuracy	Precision	Recall	ROC–AUC
CatBoost	**0.711**	**0.712**	**0.720**	**0.776**
XGBoost	0.709	0.711	0.715	0.773
Random Forest	0.676	0.676	0.690	0.724
Logistic Regression	0.649	0.658	0.639	0.709

Table 2. Test set performance comparison across models.

In words, CatBoost's performance means that when the model predicts an **outcome** will be chosen (Saved = 1), it is correct about 71% of the time (precision). It also successfully identifies about 72% of all outcomes that truly were chosen (recall), showing good sensitivity. Its ROC–AUC score of 0.776 indicates strong overall ability to distinguish between saved and not-saved outcomes across all thresholds. Although the model is capturing meaningful moral decision patterns, it also shows that a substantial amount of variation in human judgment remains unexplained. Human moral choices are highly nuanced and context-dependent, and no model can fully capture the complexity of those decisions.

Still, tree-based models overall outperformed Logistic Regression, reinforcing the importance of non-linear relationships and interactions in moral-choice prediction. CatBoost and XGBoost were nearly tied, but CatBoost had advantages in handling categorical variables and missing values natively.

5.3 Plots and Visualizations

ROC Curves & Precision–Recall Curves.

The ROC curves show that CatBoost dominates across most of the threshold range, maintaining a consistently higher True Positive Rate for a given False Positive Rate. Similarly, CatBoost exhibits the highest Average Precision and sits well above the baseline prevalence line. This indicates superior performance in identifying outcomes that will be saved, especially in the low-recall, high-precision region.

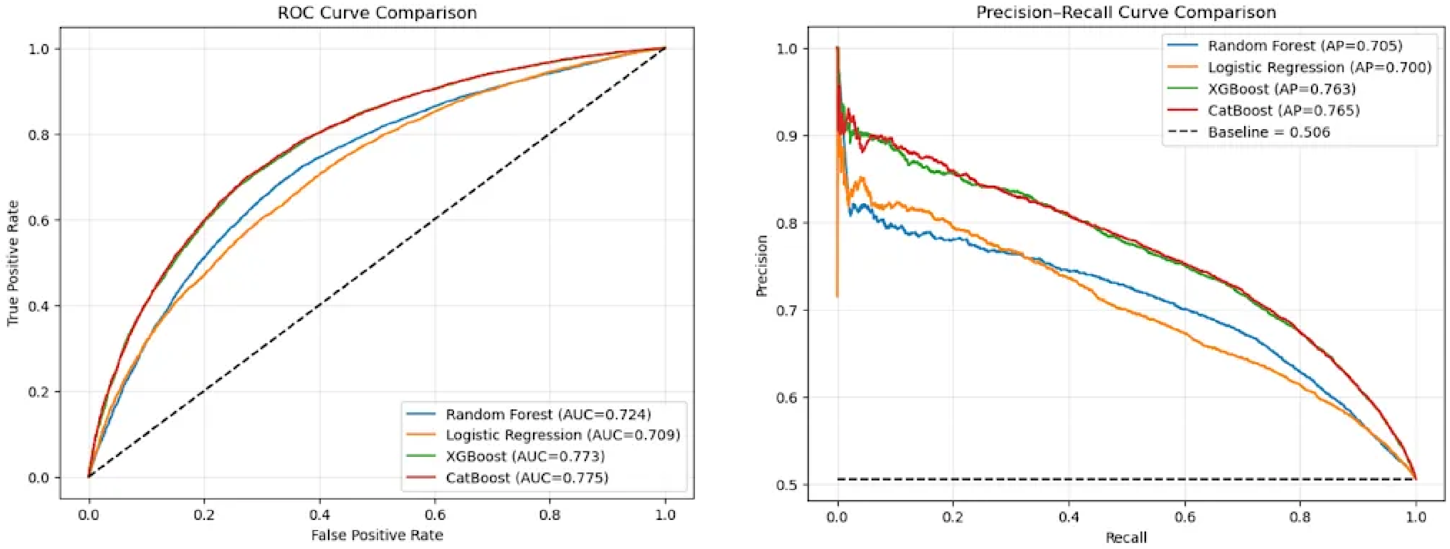


Figure 9. ROC curves (left) and Precision–Recall curves (right) for all models on the test set.

SHAP (SHapley Additive exPlanations) Values.
To interpret which features most strongly influence the model's predictions, we apply SHAP to the trained CatBoost model.

```
import shap

explainer = shap.TreeExplainer(cat_model)
shap_values = explainer.shap_values(X_test)

shap.summary_plot(shap_values, X_test)
```

In **Figure 10**, SHAP analysis reveals clear patterns in feature influence:

- **Higher likelihood of being saved:** younger individuals, women, pregnant individuals, and higher-status roles (e.g., executives, doctors).
- **Lower likelihood of being saved:** animals, elderly characters, homeless individuals, and criminals.

These findings closely align with moral psychology research, which shows that people consistently prioritize individuals who are more vulnerable (such as children or pregnant women), who hold greater perceived social value (such as doctors or executives), or who behave in norm-abiding ways. Research also highlights that people tend to extend greater moral consideration to those within their perceived "moral circle," meaning humans are prioritized over animals, and socially stigmatized groups (such as criminals or the homeless) often receive lower moral worth in ethical trade-offs. Our model's SHAP patterns align strongly with these well-documented tendencies, suggesting that the model is not inventing biases but learning real human decision heuristics embedded in the data.

However, even for highly influential features, SHAP points span both positive and negative contributions. The spread of SHAP values within each feature shows that even outcomes with the same trait are not always judged consistently. Human ethical reasoning is complex and subjective; a model can approximate but never fully replicate it.

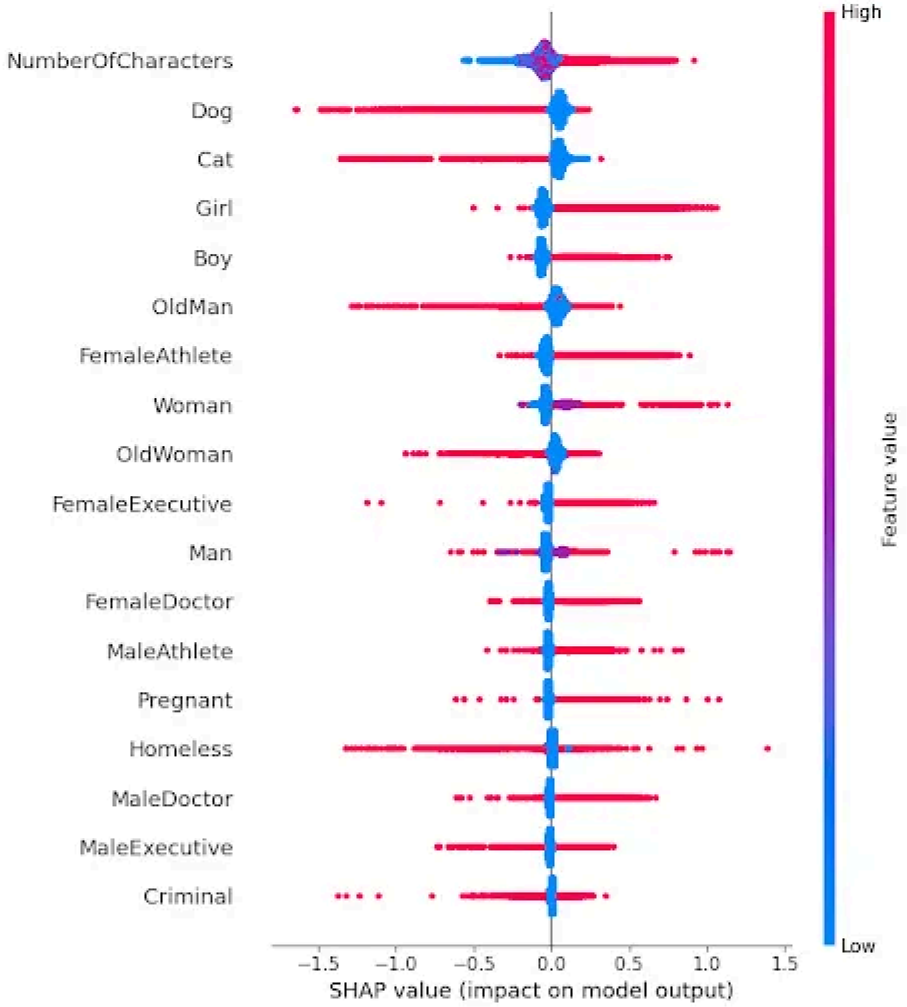


Figure 10. SHAP summary plot showing feature importance and direction of influence for the CatBoost model.

6. Discussion

6.1 Why Temporal Analysis Was Not Possible

One of the original goals of this project was to analyze temporal dynamics and investigate whether moral preferences shift over time. However, the structure of the Moral Machine dataset makes temporal analysis difficult. The dataset does not include timestamps, session identifiers with absolute ordering, or any globally consistent measure of when each judgment was made. The only ordering variable provided is `ScenarioOrder`, which reflects the sequence of scenarios shown within a single user's session, not the chronological time of data collection.

If a user decided to assess scenarios across multiple sessions, each session would reset `ScenarioOrder`, making it impossible to determine which sessions came first. Additionally, over 90% of users did not participate more than once, so any temporal conclusions based on a small subset of repeat users would not be representative of the entire population. Because participants completed varying numbers of trials per session and sessions were not associated with time stamps, there is no reliable way to reconstruct the sequence of decisions across users. Hence, assessing temporal drift was not feasible with our dataset.

As a result, we focused our analysis on cross-sectional patterns rather than longitudinal ones. This shift allowed us to study stable versus variable moral preferences across countries, scenario types, and attributes, while acknowledging that temporal trends cannot be inferred from the available data.

6.2 Business Applications

From an industry perspective, modeling moral preferences helps organizations anticipate user trust, regulatory expectations, and public acceptance of automated systems. As shown in our interpretability results, certain attributes, such as legality, youth, and perceived social value, consistently influence human decisions. These patterns can inform several business contexts.

In **marketing and consumer analytics**, insights about moral salience can guide message framing and brand positioning. Themes related to safety, family, and social contribution often evoke strong positive sentiment, and understanding these tendencies helps companies design more resonant campaigns. By associating brands with these themes, a brand can be elevated in the eyes of the public, increasing appeal.

In **insurance, risk assessment, and financial services**, human moral preferences influence perceptions of fairness, blame, and responsibility. Our results reflect strong penalties for rule-breaking behavior such as jaywalking, which parallels how consumers evaluate fairness in claims processing or pricing. Businesses can use these insights to align automated decision pipelines with customer expectations and regulatory norms.

More broadly, companies deploying **AI-driven decision-making tools** (e.g., robotics, autonomous systems, safety applications, and customer service) benefit from understanding how users perceive trade-offs. Systems that behave in ways that contradict intuitive moral norms may be seen as untrustworthy or misaligned. Incorporating moral preference modeling into system design therefore supports higher adoption, reduced ethical risk, and improved user perceptions. Overall, the business value lies in anticipating how people evaluate trade-offs and ensuring that automated systems appear to reflect values that customers and stakeholders emphasize, such as fairness and equity.

6.3 Limitations

This project has several limitations, including those highlighted in our presentation:

- **Limited temporal modeling.** The dataset lacks the temporal structure necessary to study how user preferences change over time or across repeated exposures to dilemmas. Even if we were to make temporal conclusions, they would not be representative of the entire population.
- **Uneven representation of certain scenario types.** Social status categories (e.g., executives, professionals) appear far less frequently than common

character types, potentially leading to noisier estimates of their importance.

- **Differences between experimental and real-world decision-making.** In the Moral Machine experiment, people are presented with complete information and face no penalty for taking long amounts of time for analytical thinking. In real crisis scenarios, where people must make split-second decisions, a strong status quo bias is often observed: people default to staying on course regardless of the outcomes. Because the conditions of data collection do not fully reflect real stay–swerve decisions, our model may not be directly applicable to real-world AV crash responses.
- **Outcome-level modeling rather than paired-dilemma modeling.** Our approach intentionally treated each outcome row independently rather than computing paired Δ-features, which means the model learns which outcomes are preferred, not the explicit pairwise comparison structure.
- **Subset limitations.** We used a 100,000-row sample for computational feasibility. Although diverse, it may not perfectly represent the full 40-million-row dataset. There is always the possibility that the subset does not capture the same trends that would be seen in the entire population.

Despite these constraints, the results offer meaningful insight into the structure of human moral preferences and provide a foundation for applying moral prediction models across both technical and business contexts.

7. Conclusion & Implications

In this project, we tried to understand whether human moral decisions in autonomous-vehicle dilemmas are predictable. Using a large sample of outcome-level data from MIT's Moral Machine experiment, we explored how people choose between two possible outcomes when an AV is forced to “stay” or “swerve.”

From our EDA and segmentation analysis, one of the biggest takeaways was that people do follow certain moral patterns. Across countries, respondents tended to save the larger group, younger individuals, and legal crossers, and they generally preferred to save humans over animals. Groups like doctors or executives were also more likely to be spared, while criminals or homeless individuals were saved less often. These patterns showed up clearly before modeling and later aligned with our SHAP explanations, which gave us more confidence in the consistency of these trends.

CatBoost ended up being our best-performing model, with ROC–AUC around 0.78. What surprised us most was that model performance stayed fairly stable across countries, meaning that despite cultural differences, people still showed similar decision patterns in these dilemmas. However, when we looked at different scenario types, some categories, such as Social Status or Gender dilemmas, were much harder to predict. This suggests that some moral questions are simply more uncertain or context-dependent.

Overall, we learned that even though moral decisions seem complicated, they still have enough structure that machine learning models can pick up meaningful signals. We also learned that not all moral decisions behave the same way. Some are relatively predictable, while others are inconsistent or noisy. This helped us understand where ML has strength and where human reasoning becomes too variable.

If we continued this project, we would want the dilemmas to look more like real-life emergencies. For example, people usually do not have much time to think in an accident, so adding a time limit could change how often people "default" to the stay option. Prior work suggests that many collisions could be avoided with even one additional second of decision time; limiting response time would better mimic real-world pressure. We also want to test situations where not all information is known. One idea is to show only the "stay" outcome clearly and make the "swerve" side uncertain, forcing participants to choose between guaranteed casualties versus a risky alternative with unknown consequences. These kinds of extensions would help bring the experiment closer to how people actually behave in crisis situations and could make future models more realistic and practically relevant.

8. Project Links

- **GitHub repository:** click here
- **Dataset (Moral Machine, OSF):** https://osf.io/3hvt2/

References

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

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

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 1–37.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4765–4774).

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (pp. 6638–6648).

Moral Machine Dataset (MIT Media Lab). (2016–2018). *SharedResponse.csv*. Open Science Framework. https://osf.io/3hvt2/

Written by Aileen Li

0 followers · 1 following

Edit profile

No responses yet

Aileen Li

What are your thoughts?