Learning Individual and Collective Priorities over Moral Dilemmas with the Life Jacket Dataset

Farhad Mohsin[1] , **Inwon Kang**[1] , **Pin-Yu Chen**[2] , **Francesca Rossi**[2] and **Lirong Xia**[1]
[1]Rensselaer Polytechnic Institute
[2]IBM Research

Abstract

With the increasing use of artificial intelligence (AI) in automated decision-making, machines imitating human moral behavior has become an important topic of research. The predictive power of machine learning have been used to learn cognitive models of how humans make decision while facing moral dilemmas. We collect preference data in moral dilemmas in order to learn individual preference models to represent moral priorities. We believe our newly collected dataset will be beneficial to both the AI ethics and the preference learning researchers. The dataset consists of agent preference data over specific alternatives in a particular moral dilemma (*"who gets the life jacket?"*). We test various known learning frameworks on our dataset with the goal of learning both individual and collective preferences. We find that heuristic-based lexicographic preference models (defined by a priority over features, such as in *"save women and children first"*) have accuracy comparable to more complex machine learning models in learning individual agent preferences. Finally, discuss how voting rules can be used with learned individual lexicographic preference models to predict how a group of individuals would collectively react to a moral dilemma.

1 Introduction

Humans often have to make moral judgments in various scenarios. With the rapidly growing use of artificial intelligence (AI) in various decision-making applications, this issue of making moral judgments cross into the domain of AI systems as well. For example, the often studied example of a self-driving car [Badue *et al.*, 2020] faced with an unavoidable scenario – if it must choose between harming pedestrians or its passengers, what should it do? Another such scenario is related to kidney exchange problems [Freedman *et al.*, 2020], where organ allocation protocols match donors with patients. With very high number of donors and patients and complicated inter-dependency, it becomes imperative for an algorithm to at least guide the decision. AI system designers have to encode moral values into such systems to deal with dilemmas. But for an AI system to have such moral values, it becomes important to understand how humans make such decisions in the first place. We investigate this problem of learning human models for moral decision making with a data-driven approach.

The idea of imitating how humans make moral decisions in AI system (or machines in general) has been around for a while, falling under the umbrella term: machine ethics [Wallach and Allen, 2008]. But this has recently become a topic of renewed interest with AI systems having more applications, and programming complex models being possible with high computation power. We refer to Tolmeijer *et al.* [2020] for a survey of different aspects of the machine ethics problem. Many of these works choose an ethical theory (or a combination of multiple) such as deontology [Alexander and Moore, 2021], consequentialism [Sinnott-Armstrong, 2021], virtue ethics [Hursthouse and Pettigrove, 2018], etc, and implement a model that has implicit ethical considerations for solving other problems or implement an explicit ethical agent that can make decisions in moral dilemmas.

There are two major directions here. First, the top-down approach[Wallach and Allen, 2008], encodes ethical rules chosen by the systems designer. Second, a bottom-up data-driven approach that tries to learn people's moral preferences by first gathering large amount of data and then learning predictive models. One such example is the moral machine experiment [Awad *et al.*, 2018], which frames the self-driving car's dilemma as a trolley problem and collects preference data from participants over pairs of alternative outcomes. The feature-based framework of the trolley problem and the large gathered data has allowed solutions aimed at learning ethical principles ([Noothigattu *et al.*, 2018; Kim *et al.*, 2018]) using various learning methods. The problem discussed by Awad *et al.* [2018] takes a deontological approach asking whether most surveyed people would intervene or not given a certain scenario. However, many such studies also fall into the consequentialist nature, by taking a utilitarian approach. More recently, another goal of the moral machine data has been set as helping to learn how humans make ethical decisions [Agrawal *et al.*, 2019; Awad *et al.*, 2020].

Even in this later goal of learning human cognitive model, it can be better to look at explainable models, as decisions in moral dilemma is obviously a high-stake scenario. Gigerenzer and Goldstein [1996], in their seminal work, presented how humans often make decision with minimal information using heuristics, such as a lexicographic model over the features of

alternatives. For example, the infamous accident of the early 20-th century transatlantic liner *Titanic* had a well known "Women and children first" rule when facing ship-wreckage. Rudin [2019] points out how for high-stake decision-making, interpretable models can be more preferable to black-box models. In this work, we learn such a heuristic model – a lexicographic preference model– from expressed preferences, and compare performance of said heuristic model with more complex models. We additionally discuss the problem of aggregating individual models to estimate a social model for moral preferences.

While several datasets are available for preference learning or learning-to-rank scenarios, techniques learned from them may not cross over well into the ethical domain. On the other hand, the moral machine dataset, which has been designed to enquire about moral dilemma decisions, is limited to only pairwise preferences. Thus, we felt motivated to create a new dataset that will be of use to both the preference learning and the AI ethics community. Our goal was to create a labeled dataset that lends itself to the purpose of testing existing preference models and aggregation techniques to learn ethical principles, and also of testing ethical rule/constraint-based methods that can be aligned with the collected data. While building the dataset, we also collected additional data regarding importance of features. This was done to test the learning power of heuristic-based models like lexicographic preferences [Schmitt *et al.*, 2006], which are an important heuristic-based decision-making methods.

A major contribution of this work is a new dataset (the Life Jacket dataset), which considers a moral dilemma that can be seen as a significant variation of the trolley problem that allows rankings over more than two alternatives. We also collect ground truth preferences over features of alternatives, that we then use to test learning methods over the dataset. Details of our dataset can be found in Section 3.

Experimentally, we compare the performance of various preference models, such as classification algorithms, utility based models, and heuristic models like lexicographic preferences on the new dataset. We see that, for our dataset, the easily explainable lexicographic preference model performs comparably (Section 5) to more complex black box type models including neural networks.

2 Related Work

In this section, we discuss related preference datasets both in ethical and non-ethical domains. Also, we discuss preference learning frameworks [Cohen *et al.*, 1999], as they are relevant for our modeling of preferences in moral dilemma.

2.1 Existing Preference/Ethics Datasets

There are numerous datasets that consider preference over different alternatives in the morality domain. For example, Eriksson *et al.* [2021] consider the 'morality' section from the EVS dataset[EVS, 2011] to find the correlation between perceived morality and commnality of different actions. However, this dataset contains people's preference for or against certain pre-set actions, which makes it hard generalize to other scenarios. Awad *et al.* [2020] consider people's preferences on cutting lines given different situations, but again the problem is posed as a yes/no question of type "is this moral?" rather than a preference over alternatives.

LETOR [Qin *et al.*, 2010; Qin and Liu, 2013] is a dataset of documents rankings, where the position in the ranking is related to the relevance of the document to a particular query often used as a learning-to-rank benchmark. The rankings are actually generated by algorithms, whereas the human expert contributes only with a binary label for the documents as relevant or non-relevant. **Sushi Preference Dataset [Kamishima, 2003]** has different flavors of sushi as alternatives, from which individuals rank their top 10. Also, the number of outcomes is limited to 100 types of sushi, whereas in a moral dilemma scenario we could get many more possible alternatives. **The Netflix Prize Dataset [Bennett *et al.*, 2007]** comprises of movie ratings provided by 17,700 agents. On average, each agent rates about 200 movies, on a scale of $0-5$. The use of a limited scale means that, in pairwise preferences, many of the movies are in a tie (they have the same rating). In a moral dilemma scenario, ties are not welcome and coarse recommendations are less preferred than strict pairwise preferences. Moreover, none of these datasets can be considered for high stake preferences, like in ethical scenarios. So, preference models working on these may not even transfer well to preferences in moral dilemmas.

The Moral Machine Dataset [Awad *et al.*, 2018] comprises millions of pairwise preferences over trolley problem scenarios involving self-driving cars and is probably the best known tool to study human preferences in moral dilemmas (for example, shown by [Noothigattu *et al.*, 2018; Kim *et al.*, 2018; Wiedeman *et al.*, 2020]). The problem statement is for a self-driving car that cannot avoid an accident and needs to decide between alternative directions. Individuals are asked to choose and therefore to give a moral judgement on what action is more acceptable from an ethical point of view. The dataset collects a huge number of such judgments from a large population. All judgements are related to pairwise comparisons, which is a severe limitation. Additionally, while the full dataset is large, on average the number of pairwise preference data for each agent is rather low.

2.2 Modeling and Aggregating Preferences

A common way to collect and study preferences in an ethical domain is to model the problem as a decision between two alternatives, each including some problem features. This can be thought of as a classification problem, where the classifier's decision is to either 'intervene' or 'not'. Several pieces of work, such as [Shaw *et al.*, 2018] and [Wiedeman *et al.*, 2020], tackle this problem from a bottom-up statistical learning-based approach. This leads to decisions that –while consistent with the data– are difficult to explain.

On the other hand, moral dilemmas can also be thought of as general preference problems, where the preference can be over more than two outcomes [Rossi, 2016]. Common ways of expressing preferences include using utility functions, or structured preference models like CP-nets [Boutilier *et al.*, 2004], or lexicographic preference models [Schmitt *et al.*, 2006]. Utility functions usually assume some quantitative score/valuation for each alternative, which determines the preferences.

	LETOR 4.0	Sushi	Netflix	Moral Machine	Life Jacket
Ethics Domain	N	N	N	Y	Y
Non-binary comparisons	N*	Y	Y	N	Y
Feature importance reported	N	N	N	N	Y
Agent features reported	N	N	N	Y	Y
No. of agents	N/A	5,025	17770	∼2,600,000	673
No. of pairwise preferences	N/A	226,125	$\sim 10^{10}$	$\sim 4 \times 10^7$	20,190
Pairwise preference per agent	N/A	45	∼20,000	∼14	30
No. of total alternatives	10000	100	17,700	>100,000	5000

Table 1: Comparison of the Life Jacket dataset to other publicly available preference datasets

[Noothigattu *et al.*, 2018] uses random utility models with features and deploys a voting mechanism to aggregate individual preferences. [Kim *et al.*, 2018] makes a further assumption that all individuals are sampled from some common distribution that determines their utility scores, using a hierarchical Bayesian model. [Awad *et al.*, 2020] uses CP-nets to solve a moral dilemma of when it is morally acceptable to break a rule. In this paper, we study the viability of using simple lexicographic models to model ethical preferences, compared to more complex classification algorithms. Multiple learning algorithms have been proposed (e.g., [Booth *et al.*, 2010; Yaman *et al.*, 2011]) to learn individual moral preferences. These are discussed in more detail in Sections 4 and 5.

For a preference model to be socially acceptable, it usually needs to reflect the principles of a society rather than an individual. In that regard, methods learning individual models need to be aggregated to indicate a social principle. Voting rules are one of the most popular techniques for aggregating individual preferences and multiple voting rules with different properties (some popular ones are plurality, Borda, Copeland) have been developed [Brandt *et al.*, 2016]. When facing a new scenario, [Noothigattu *et al.*, 2018] propose applying voting rules on the predictions of individual preferences to get a group prediction [Kahng *et al.*, 2019]. Related work are done by [Li and Kazimipour, 2018] and [Lang *et al.*, 2012], who discuss the general problem of aggregating lexicographic preference trees and aggregating agent preferences when said preferences are lexicographic in nature.

3 The Life Jacket Dataset

The moral dilemma scenario we focus on is the following: *Suppose an airplane is about to crash and there is only one rescue jacket left but more than one person on the plane. Whom would you (an external observer) prefer to give the jacket to?*[1]

It can be seen as a generalization of the trolley problem where there are more than two alternatives to consider. Our data collection method had individuals presented with multiple alternatives (2, 3 or 4 alternatives at a time) composed of 7 features - age, gender, health, number of dependents, survival chance with and without the life jacket - and asked to rank them according to whom they would prefer to give the life jacket to. What changes among the various scenarios is the list of alternatives that are presented to the survey participants. By randomly changing the alternatives in each scenario presented, we aimed to ensure that we cover as much as possible in the feature space.

Feature	Possible values
age	$5 - 72$
gender	male, female
health condition	in great health, small health problems, moderate health problems, terminally ill(< 3 years left)
income level	low,mid,high
number of dependents	$0 - 5$
survival chance without jacket (%)	$[1, 40]$
survival chance increase with jacket (%)	$[1, 50]$

Table 2: Domain for each feature variable in our survey

3.1 Data Collection

Using Amazon Mechanical Turk, we collected data from 673 participants[2] (hereon called agents). Agents are asked to read a short description for the moral dilemma, where an aircraft carrying several people is about to crash and there is exactly one life jacket. Before the main survey stared, we also asked for some basic (age, gender, education level) demographic information from the agents. Each agent was faced with 17 scenarios in total (12 with two alternatives, 4 with three alternatives, 1 with four alternatives), and for each scenario he/she was asked to rank the alternatives in terms of who they would prefer to save. After the preference questions, we also asked the agents to mark their perceived importance of each feature of the alternatives, as well as a text input to explain their reasoning behind the choices. We present additional our survey workflow in Figure 1. A sample scenario from an actual survey is shown in Figure 2. Additional details for the dataset using the methodology of datasheets for datasets Gebru *et al.* [2018] is provided in Appendix A.

3.2 Survey Design

To ensure that the features in our dataset would cover a sufficient amount of the feature space, we generated 5000 alterna-

[1]We plan to release the dataset for public use after publishing of the paper.

[2]We had 700 participants but we discarded 27 because of noisy data.

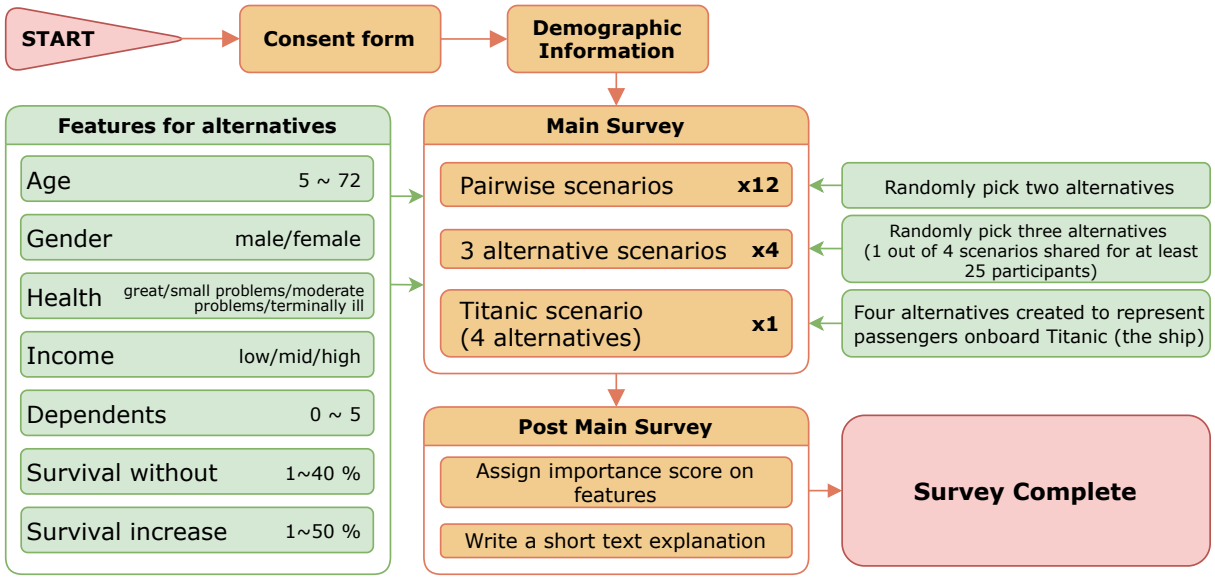

Figure 1: A diagram of the survey workflow

tives with randomly chosen feature values. In order to avoid alternatives that were unrealistic, such as a 9 year old child supporting any number of dependents, we had some rules for the generation process to avoid such combinations. The features were chosen from a domain shown in Table 2. A plot of the distribution of features is available in Appendix A. For each agent, we then randomly generated twelve scenarios with two alternatives and three scenarios with three alternatives. One of the three-alternative scneario an agent faces is also faced by around 24 other agents. Each agent is also provided with another four-alternative scenario that is common to all agents. We call this the Titanic scenario . These common scenarios were generated with the goal of being presented to disjoint sets of 25 agents to create common data for testing preference aggregation schemes. For this, we actually create representative alternatives presenting passengers on the Titanic Kaggle [2012]. We chose four passengers that were cluster centers in terms of their features that (gender, income, health, dependents) and translated those features into our respective feature space. By using a real-life example, we wanted to check how our learned models would predict, as compared to the 'women and children first' rule.

4 Preliminaries for Analysis

Assume we have n agents and a set of alternatives $\mathcal{A} = \{a_1, \ldots, a_m\}$. Each alternative can be defined by a feature vector with variables $(X_1, \ldots, X_k)$. Abusing notation, we also call the feature vector for alternative a_1 as $a_1 = \langle x_1^1, \ldots, x_1^k \rangle$. For each feature variable X_i, we assume $\mathrm{Dom}(X_i)$ is the domain for the variable, and thus each $a_i \in \mathrm{Dom}(X_1) \times \ldots \times \mathrm{Dom}(X_k)$.

Each agent will have preferences over the alternatives. Preferences can be considered as weak ordering over $\mathcal{A}$, where weak ordering is an ordering over the full set with the possibility of ties. Let $\mathcal{B}(\mathcal{A})$ be the set of all weak orders over $\mathcal{A}$. Thus, with the n agents, we get a preference profile $\in \mathcal{B}(\mathcal{A})^n$. Now, these preferences can be broken down into pairwise preferences between the alternatives as well, for example for any two alternatives a_i, a_ℓ, and an agent j, three things can happen - a_j may be more preferred by j, $a_i \succ_j a_\ell$; a_ℓ may be more preferred $a_\ell \succ_j a_i$; or they may be equally preferred $a_i \sim_j a_\ell$.

Looking at preferences as a collection of pairwise comparisons, a classifier can be used to model ethical preference. For every pair of alternatives, consider the input feature to be the two alternative features, say (a_i, a_ℓ), and the decision is which one is more preferred. A classifier $f : (a_i, a_\ell) \to \{0, 1\}$ is a mapping from a pair of two alternatives to a binary preference. Using this approach, we will ignore the tied data points for the classifiers. In this paper, we make use of traditional classifiers like logistic regression and neural networks. Our experimental setup only makes use of 7 alternative features. Due to the low feature size and low dataset size, we limit our experiments to shallow densely connected neural networks of 1 and 2 layers with ReLU activation function.

A lexicographic preference model (LPM) is defined using a total order over the set of feature variables itself and a preference table, defining for each feature. Assume such an order to be $\mathcal{L} = X_{\ell_1} \succ X_{\ell_2} \succ \ldots \succ X_{\ell_k}$. Here, the variable that is ranked highest will be considered the most important. Now, given two alternatives a_i, a_ℓ and a order over features, $\mathcal{L}$, to find preference we find the highest ranked feature that is different between the two. Suppose, said variable is X_j. So the preference between a_i, a_ℓ will depend solely on the feature values x_i^j, x_ℓ^j. We assume discrete feature variable domains for lexicographic preference models, and to keep the model simple, in accordance with existing literature, we assume that the preference over values of a single variable would be monotonic in nature. For example, in a Trolley problem like scenario, if age is the most important feature, the preference would be

	Person 1	Person 2	Person 3
:age	32	18	8
:health	small health problems	in great health	terminally ill (less than 3 years left)
:gender	female	female	male
:income	high	low	mid
:#dependents	1	2	0

Please assign each option a score in the table below

These scores are relative to each other and are meant to express the ranking between them.
The more you think this option should be chosen, the higher the score should be.
example: A score of A: 10, B: 8, C: 6 would mean you want to choose option A the most, then option B and finally option C.

	Survival Chance			
	Person 1	Person 2	Person 3	Score
Option A: give the jacket to Person 1	58%	25%	3%	
Option B: give the jacket to Person 2	11%	70%	3%	
Option C: give the jacket to Person 3	11%	25%	48%	

Figure 2: A sample scenario from the actual survey

"save the young" or "save the older", but we avoid preferences like "save middle aged people first, then the children, then the older people". Simple lexicographic preferences are in fact a special case of lexicographic preference trees. The simple lexicographic preference model that we focus on in the analysis section is called an *Unconditional Preference-Unconditional Importance (UP-UI)* tree [Booth *et al.*, 2010]. We show an example lexicographic preference model (LPM) in Figure 3 to illustrate the model. Here, we have the order of features $\mathcal{L} = Age \succ Dependents \succ \ldots \succ Gender \succ Income$. Also, for each feature, we have a preference over the features, e.g., for Age, the preference under this LPM is to save children first, then middle-aged people and finally older people. If the alternatives are a middle aged male with 3 dependents and a middle aged female with 2 dependents, the former would be preferred by this LPM because number of dependents is the highest ranked feature where the values disagree.

A *voting rule* $r : \mathcal{B}(\mathcal{A})^n \mapsto \mathcal{A}$ is a function that maps a preference profile to a single winner. We further define two popular voting rules, plurality and Borda. Plurality and Borda are both scoring rules, which means each alternative gets a score according to which position each agent has placed them overall. The alternative with maximum score wins. The score vector with m alternatives for plurality is $\langle 1, 0, \ldots, 0\rangle$ and for Borda, it is $\langle m-1, m-2, \ldots, 1, 0\rangle$. In case of ties, there are various ways of tie-breaking, but for simplicity, we just consider lexicographic tie-breaking during our experiments, where ties are broken in an arbitrary but fixed order of features.

Kendall's Tau (KT) correlation coefficient is a measure of rank correlation . A positive KT coefficient means that the two rankings are somewhat similarly aligned, with complete alignment at $+1$. Similarly a negative KT coefficient means dissimilarity with exact opposite rankings giving the value of -1.

5 Modeling Priorities in Moral Dilemma

Much work has been done in using machine learning techniques in modeling priorities in moral dilemmas (e.g., by Awad *et al.* [2020]; Noothigattu *et al.* [2018]; Wiedeman *et al.* [2020]). Most of these work depends on casting the moral dilemma problem as a binary classification problem. The task is transformed into a binary classification task, a choice between two alternatives (in particular, work based on the moral machine dataset works with the two alternatives - intervene or not). However, in our new dataset, and in many realistic scenarios, the goal may be to get a full linear order over alternatives. Additionally, using traditional classifiers often lead to black-box type models or predicting utility values for alternatives. Both of these are undesirable when the task is to model how humans make decisions in moral dilemma. An interpretable or explainable model makes more sense to model agents' priorities regarding moral issues. We take motivation from the famous work by Gigerenzer and Goldstein [1996], which proposed that humans often make such decisions in a "fast and frugal way", by making use of some sort of heuristic. Thus, when analyzing our Life Jacket dataset, we compare traditional learning algorithms and lexicographic preference models, which is such a heuristic-based model.

Learning from Life Jacket Dataset

Hypothesis 1.1 Individuals have lexicographic preferences (e.g. save children first, then women, then the rich etc.) when making moral decisions

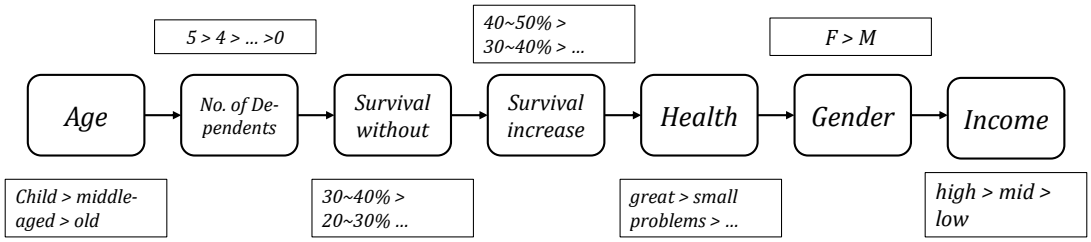

Figure 3: Example lexicographic preference in the Life Jacket domain

	Lex-learning	Logistic Regression	Decision Trees	1-Hidden layer NN	2-Hidden Layer NN
Training Accuracy	0.78	0.79	0.82	0.84	0.90
Testing Accuracy	0.62	0.64	0.64	0.65	0.64

Table 3: Mean accuracy comparison between Lex learning and other learning techniques for individual models for each agent

	Plurality	Borda
Lex Learning	0.52	0.59
Decision Trees	0.34	0.55
Logistic Regression	0.62	0.38
1-Hidden Layer NN	0.38	0.62
2-Hidden Layer NN	0.38	0.62

Table 4: Mean accuracy comparison between Lex learning and other learning techniques for aggregate decision prediction

Brute force searching through all possible lexicographic preference models (LPM) over the set of features, we find the best LPM for each agent. Then we break down each ranking and try to predict the pairwise preference for each pair. The "best lexicographic preference" for each agent gives us an average 79.8% accuracy over all agents. This indicates that our hypothesis is possibly true because the best lexicographic preference is far more effective than a random binary classifier.

Hypothesis 1.2 We can learn to predict lexicographic preferences for individual agents

Even if agents have lexicographic preferences, real-word data is rarely noise-free, and we would get inconsistent examples with the underlying ground truth model. It has been shown, that in presence of inconsistent samples, finding the best LPM is an NP-complete problem (Proposition 7 in [Booth *et al.*, 2010], our assumed model is the UP-UI model mentioned in that paper). For this, it is unlikely to get much better performance than a greedy learning algorithm [Yaman *et al.*, 2011].

We implement Algorithm 1 (which we will refer to as **Lex Learning**) to learn LPMs from individual agents without using a brute force search. This is a variation of the greedy lexicographic preference-learning algorithm (Algorithm 1 in [Yaman *et al.*, 2011]) to learn individual lexicographic preference for each agent. To account for noisy data, our variation takes a randomized-greedy approach instead of the regular greedy algorithm, as shown in Algorithm in the Appendix. The algorithm learns priorities over the features and the preference table for each feature. Since we are assuming monotonic preferences for each feature, it is not necessary to learn a full preference table, rather the monotonic direction for the preference. For example, in our example in Figure 3, the preference for Age was $Child \succ middle-aged \succ old$. Thus, smaller values are more preferred. Under such assumption of monotonicity, the preferences for each feature can be defined using a binary direction variable, which we call dir. At any time, with set of observations P, for a feature X and either direction value of dir, we can define variable-direction consistency. For example, w.l.o.g. assume dir is increasing, then the tuple (X, dir) is consistent with a pairwise preference p between alternatives a_i, a_ℓ if $x_i > x_\ell \iff a_i \succ a_\ell$. We quantify it with indicator random variables in the following way.

$$\text{consistent}(X, P, dir) = \frac{\sum_{p \in P} \mathbf{1}_{(X,dir) \text{ consistent with p}}}{\sum_{p \in P} \mathbf{1}_{x_i \neq x_\ell}}$$

Then Algorithm 1 greedily chooses the feature and direction that is most consistent with the available set of observations best and repeatedly adding new features to create a complete LPM. Because of the randomized nature of Algorithm in appendix, we repeatedly run it a number of times, and consider the one with highest training set accuracy.

Training on the whole dataset, we get a training set accuracy of 78.2%. However, doing a 5-fold cross validation, we get a cross validation accuracy of 62.2% for each agent. This indicates that more samples may be needed to learn individual accuracy correctly. We explore this in the synthetic data experiments in Appendix A.

Hypothesis 2: People are inconsistent in ranking with reported feature importance

We make this hypothesis that people do not give completely accurate rankings in terms of their own reported feature importance scores. And we test it by matching the "best lexicographic preference" to their reported importance scores and computing the ranking distance, using Kendall's Tau (KT) correlation coefficient. Since we do not have any information

Algorithm 1 Pseudo-code for Lexicographic Preference Learning (Lex Learning)

1: **Inputs:** Set of features V, and set of observations P of pairwise preferences
2: **Initialization:** Empty lexicographic preference model $\mathcal{L}$
3: **while** V and P both non-empty **do**
4: For each $X \in V$, for each direction dir for X, compute

$$\text{consistent}(X, P, dir) = \frac{\sum_{p \in P} \mathbf{1}_{(X,dir) \text{ consistent with p}}}{\sum_{p \in P} \mathbf{1}_{x_i \neq x_\ell}}$$

5: Sample a (X, dir) with probability proportional to $\text{consistent}(X, P, dir)$ and add (X, dir) to $\mathcal{L}$
6: Remove X from V
7: Remove all $p \in P$ consistent with (X, dir) from P
8: **end while**
9: **Output:** Learned lexicographic preference, $\mathcal{L}$

about direction for the feature importance scores, we discard that information from the found LPM and use a similarity metric depending only on the order or ranking.

The average of KT correlation coefficient between the LPM and order of reported feature importance for all agents is 0.37. While positive correlation means that they are somewhat consistent, this coefficient is still low for 7 features, which means that the reporting of importance features is not completely consistent with the actual preferences expressed in scenarios. For example, in Figure 4, we see that gender and income has low average importance scores, indicating that these features are considered less important than the others. However, in the best lexicographic preferences found by using brute force, many agents considered income and gender before the two survival chances in lexicographic order. We also notice that survival chances are probably something agents overlooked when actually observing the alternative even though separately they thought they are important features for this dilemma.

To get another notion as to whether the two forms of information are somewhat inconsistent, we transform the feature importance scores into lexicographic preferences. Now, since for each feature, the local preference can be either in favor of decreasing or increasing (e.g. save the young first vs save the old first), we try all possible such lexicographic preferences. Then we calculate the accuracy that this model gets on the complete training data. This leads to an average of 65.2% accuracy in predicting the judgment of pairwise comparisons, which is much worse than what we have for the best lexicographic preference found using brute force search.

This set of experiments indicate that while reporting importance scores may still give a good idea about priorities regarding moral dilemmas, a better way to learn individual preference model is likely by introducing moral dilemma scenarios and learning a model from them.

Hypothesis 3: We can predict aggregate preferences despite low individual accuracy

Even though our cross-validation accuracy is not high for individual agents, we predict group decisions by applying voting rules to the predictions. We have a total of 29 aggregate scenarios. The accuracy for correctly predicting the aggregated moral decision for different voting rules and different learning algorithms is given in Table 4. Here, a correct prediction means that applying the voting rule on the ground truth and on the prediction (when training on everything else) gives the same result. Here also we see that the Lex Learning algorithm performs comparably to other learning algorithms.

While the results may seem poor, this is not unexpected because individual prediction accuracy was low as well. Also, as our aggregate test cases consider 3 or 4 scenarios, the accuracy is considerably better than random guesses. Interestingly enough 1-layer NN and 2-layer NN's give the exact same prediction in all test-cases for both voting rules, thus the equal prediction scores. It appears that predicting Borda winner is easier for most models.

Also, for the real-life inspired Titanic scenario, the aggregated prediction is correct all models for both voting rules. This may be because we have a high number of agents for this scenario, the noise is reduced in aggregation and thus we got a better aggregate prediction.

6 Discussion and Future Work

Learning preference models for humans is a sensitive topic, as it leads to requirements like interpretability. Although we should never have complete reliance on AI systems to make such choices, we think that the presence of some AI system that correctly aggregates the population preference can be helpful in that it can aid other humans make a difficult decision.

Analysis on our dataset shows how heuristic based lexicographic models perform comparably to other more complex models. A possible reason for this is that humans can also make moral judgments based on some sort of heuristics, rather than computing some sort of latent utility for each outcome in such moral dilemmas.

We also show how even in presence of noise in individual preference models, aggregate predictions is more likely to be accurate and thus collective priorities can be correctly predicted. Due to the problem of gathering a large dataset for these scenarios, we saw that none of the models performed really well. However, synthetic experiments indicate that the models might work well in aggregate with a higher number of agents even with high individual-level noise.

For future work, we would like to increase the size of the dataset, with input from a diverse population. While a set of participants on the MTurk website does not constitute a representative sample of any population, having participants from different locations would further help to remove some bias in the dataset. We discuss some issues regarding this bias in Appendix A.

Finally, while we focus on learning heuristic type of rules in a data-driven way, an interesting way to do this would be to actually have hybrid type models that is a mixture of data-driven models and ethical theories.

References

Mayank Agrawal, Joshua C Peterson, and Thomas L Griffiths. Using machine learning to guide cognitive modeling: A case study in moral reasoning. In *The 41st Annual Conference of the Cognitive Science Society*, 2019.

Larry Alexander and Michael Moore. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum, and Max Kleiman-Weiner. When is it morally acceptable to break the rules? a preference-based approach. In *12th Multidisciplinary Workshop on Advances in Preference Handling (MPREF 2020)*, 2020.

Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius Brito Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago Meireles Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, page 113816, 2020.

James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. Citeseer, 2007.

Richard Booth, Yann Chevaleyre, Jérôme Lang, Jérôme Mengin, and Chattrakul Sombattheera. Learning conditionally lexicographic preference relations. In *ECAI*, volume 10, pages 269–274, 2010.

Craig Boutilier, Ronen I Brafman, Carmel Domshlak, Holger H Hoos, and David Poole. Cp-nets: A tool for representing and reasoning withconditional ceteris paribus preference statements. *Journal of artificial intelligence research*, 21:135–191, 2004.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

William W Cohen, Robert E Schapire, and Yoram Singer. Learning to order things. *Journal of artificial intelligence research*, 10:243–270, 1999.

K. Eriksson, I. Vartanova, and P. Ornstein. The common-is-moral association is stronger among less religious people. *Humanities and Social Sciences Communications*, 8(109), 2021.

EVS. EVS - European Values Study 1999 - integrated dataset, data file version 3.0.0, 2011.

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

Rosalind Hursthouse and Glen Pettigrove. Virtue Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition, 2018.

Kaggle. Titanic - machine learning from disaster, 2012.

Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel Procaccia, and Christos-Alexandros Psomas. Statistical foundations of virtual democracy. In *International Conference on Machine Learning*, pages 3173–3182. PMLR, 2019.

Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 583–588, 2003.

Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B Tenenbaum, and Iyad Rahwan. A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 197–203, 2018.

Jérôme Lang, Jérôme Mengin, and Lirong Xia. Aggregating conditionally lexicographic preferences on multi-issue domains. In *International Conference on Principles and Practice of Constraint Programming*, pages 973–987. Springer, 2012.

Minyi Li and Borhan Kazimipour. An efficient algorithm to compute distance between lexicographic preference trees. In *IJCAI*, pages 1898–1904, 2018.

Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Tao Qin and Tie-Yan Liu. Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

Francesca Rossi. Moral preferences. In *The 10th Workshop on Advances in Preference Handling (MPREF)*, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Michael Schmitt, Laura Martignon, and Dana Ron. On the complexity of learning lexicographic strategies. *Journal of Machine Learning Research*, 7(1), 2006.

Nolan P Shaw, Andreas Stöckel, Ryan W Orr, Thomas F Lidbetter, and Robin Cohen. Towards provably moral ai agents in bottom-up learning frameworks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–277, 2018.

Walter Sinnott-Armstrong. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: a survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38, 2020.

Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

Christopher Wiedeman, Ge Wang, and Uwe Kruger. Modeling of moral decisions with deep learning. *Visual Computing for Industry, Biomedicine, and Art*, 3(1):1–14, 2020.

Fusun Yaman, Thomas J Walsh, Michael L Littman, et al. Democratic approximation of lexicographic preference models. *Artificial intelligence*, 175(7-8):1290–1307, 2011.

A Dataset Description

A.1 Description

• *Data instances:* Each data instance is an agent, i, the agent description, a set of alternatives $\mathcal{A}$ (candidates for the life jacket, defined in terms of problem's features), and the agent's ranking over these alternatives. The set of alternatives together is sometimes referred to as a scenario in the paper, since this set of alternatives is what creates a moral dilemma-type scenario for the agent to give preference on.
• *Nature of instances:* Each agent is given 17 sets of alternatives and is asked to rank each set separately in terms of who they would want to save first. 12 instances had two alternatives, 4 had three alternatives, and 1 had four alternatives.
• *Number of instances:* We have data for 673 agents in the dataset, so 11,441 instances in total.
• *What data does each instance consist of?* Each agent is identified by a unique anonymized agent ID. The features collected for the agents are age, gender, and education (agent feature information submission was optional, hence some instances lack this info).

The alternatives were created by randomly choosing specific feature values to describe a person, see below:

- Age;
- Gender;
- Health condition - Rather than specific conditions, we have the broad categories –"completely healthy", "slight health problems", "moderate health problems", "terminally ill";
- Income level - Again, rather than specific income amounts, we give the categories "high", "medium", "low", keeping the interpretation of these categories up to the agent;
- Number of dependents - Number of people (explained to be parents, children or other wards) dependent on this person;
- Survival chance if given life jacket;
- Survival chance without life jacket - These last two bits of information were presented as probabilities given that a person receives (or does not receive) a life jacket. Thus, making a choice (e.g. giving the life jacket to person 1) seems like picking an option that gives everyone specific survival chances. So, the final choice is akin to a choice over lotteries.

Table 2 shows the considered domain for each feature.
• *Data for preference aggregation:* In addition to learning individual preference models, we are also interested in aggregated preferences which can indicate a social preference model in a moral dilemma. For this purpose, we ensure that some common scenarios are presented to different agents so that preference aggregation schemes in moral dilemmas can also be tested. For this purpose, one common scenario was presented to all agents, giving us 673 agent preferences for the same set of alternatives. Additionally, we divided all the agents into disjoint sets of around 25 agents, and each disjoint set had a common three-alternative scenario given to the agents of that group. Thus, we have one common scenario (the Titanic scenario in Table 5, explained further in Section A.2) with 673 respondents. And 28 more scenarios, for each of which we have a group of around 25 respondents. With these 29 scenarios, we can test preference aggregation schemes.

Features	age	gender	income	health	dependents
Person 1	21	male	low	great	0
Person 2	32	male	low	great	0
Person 3	52	female	high	great	1
Person 4	5	female	high	great	0

Table 5: Alternatives for the Titanic scenario

• *Feature importance data:* For each agent, in addition to the 17 instances, we collected feature importance scores. So, given the features that define an alternative, agents give a score between 0-10 for each feature in terms of how important they think that feature is.

We believe the feature importance scores give us additional insight into moral preferences and is an important part of the dataset that will help further analysis. It can even help measure accuracy for some heuristic based models that depend on importance and inter-dependency of features. These scores also served as our ground truth, as we could compare the model generated weights with the self-reported scores to check the consistency of the participants' input. The central statistics for reported importance scores are presented in Figure 4. It seems from this that age and health are considered highly important, with survival chance having next high importance.

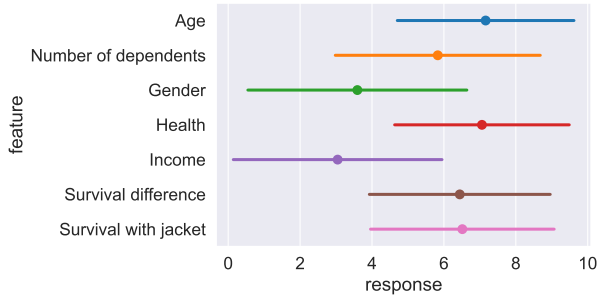

Figure 4: A scatter plot of the mean user reported scores with standard deviation

• *Recommended usage and evaluation measure:* As mentioned, we have 17 instances per agent. While this may seem like a small size to learn preference models, breaking the preferences over three or four alternatives for pairwise comparisons gives a total of 30 pairwise comparison per agent. We use the pairwise preference data to learn individual models in our experiments.

For the aggregation data, we treat the pairwise comparison from the common scenarios as test data and train on everything else. Then we run aggregation schemes like voting rules on both ground truth labels and predicted preferences and check error for aggregated decision prediction.

A.2 Data Collection Process

• *How was data collected?* Data was collected through running an IRB approved survey on Amazon Mechanical Turk (MTurk). We recruited 700 Turkers, who participated between October and December of 2020. Each individual was paid 0.85$ for their input. The data collection workflow is shown in Figure 1. Here is the task introduction the Turkers saw in MTurk:

You'll be given a moral dilemma about choosing who to save in a life-threatening scenarios. You'll be given 17 scenarios in total. You'll also be asked importance scores of different aspects in such dilemma. Completing the survey should take 5-7 minutes.

Given the 5-7 minute completion time and the payment, we estimate an hourly wage between 7.7 to 10.2 USD.

• *How were the instances generated?* We created a set of 5000 alternatives using the features in Table 2 beforehand in order to create scenarios for an agent to express their preference on. To make the data more realistic, some rules were enforced on the combinations of features while creating the set of all alternatives to rule out less likely combinations of features. For example, children under the age of 15 can not have dependents and can not have 'high' income; people over the age of 65 cannot have more than 3 dependents. Other than the cases falling under these rules, each feature value was sampled uniformly at random from its domain.

For each agent, we then randomly generated twelve scenarios with two alternatives and three scenarios with three alternatives. Each agent is also provided with another three-alternative scenario which is faced by roughly 24 other agents. These common scenarios were generated with the goal of being presented to disjoint sets of 25 agents to create data for preference aggregation, as mentioned before.

In addition, every agent also received a special scenario with four alternatives, which is given on Table 5. This scenario is based on a real life example, the Titanic Incident, and the four alternatives are taken. For this Titanic scenario, we emulate real world data from the passengers of the cruiser Titanic. We were interested to see how an example from this actual incident would be received by the agents. Using available information of passengers aboard the ship, we used k-means clustering with $k = 4$ to find the representative passengers of different demographics. Then we chose relevant features that could be translated over to ours, such as seat class (income level), age, passengers travelling with them (dependents). The chosen passengers for our experiments can be seen in table 2. For features that we used in our experiments but that are not available (i.e. health condition, survival with jacket, survival without jacket), we set equal values for them, so that the agents' choices would not be affected by those values. All of them are given the same survival chance (0% for without life-jacket and 32% for with life-jacket). With these four representative passengers defined using the representative features, we create what we call the the Titanic scenario that is presented to every agent.

• *Does the dataset contain all possible instances?* For a moral dilemma like this, the domain for all possible instances is infinite, so obviously it is not possible to get preferences over all possible alternatives. Which is why we randomly sample a feasible number of scenarios per agent for them to judge. We believe that completely random and differentiated (in case of Titanic) sets of alternatives are better proxy for real world alternatives. This is one difference in design from the moral machine dataset, where they focused more on how each feature affected agent decisions. So most of the scenarios presented there were of type male-vs-female, old-vs-young etc. We designed more random scenarios in hope of capturing the interdependence of the various feature in any model learned from this data. The distribution of feature values in our generated alternatives can be seen in Figure 5. In this figure, it can be seen that some values seem to have a concentration, such as 0 dependents or survival chance higher than 40%. This was due to the restrictions we had in the generation process, where we did not allow younger people to have many dependents and set a survival chance of at least 41% with the jacket to make sure that the value would always be greater than the survival chance without the jacket (which is capped at 40%). The focus as the lower ages is also a result of our restrictions, because we allowed for more duplicate values in fields for lower ages than the higher ages (for example, since both 5 and 7 year olds cannot have dependents, we allowed them to have duplicate values for that feature), we ended up with more alternatives with younger ages.

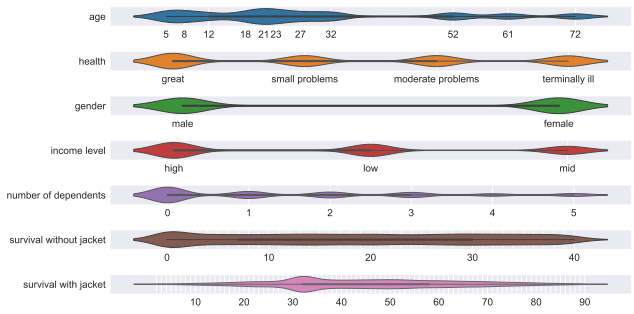

Figure 5: A density plot on the distribution of feature values in our generated instances

• *Is dataset a sample of some population?* As the dataset was collected from Amazon Mechanical Turk and no geographic limitation was placed, this dataset should not be considered as a sample for any specific population. Based on the demographic information gathered from participants, about two-thirds of the participants identified as male, where the rest identfied as female. Both the mode and median age group is $30 - 39$ years old. The percentage of participants who completed college is 28%, with the rest reporting high school or middle school as their highest level of education.

A.3 Ethical Implications

Since our dataset is concerned with ethical principles and moral dilemmas, there could be ethical implications. We want to reinforce the fact that the goal of our collected data was to give a tool for learning aggregate moral priorities and testing such learning and aggregation techniques. Any model learned from this dataset should not be deployed anywhere as an "ethical AI model". Because, as we mention above, this dataset is not representative of any population and it was not

collected to behave that way.

On the other hand, the dataset has been completely anonymized with no identifying feature of the participants collected. So, there is no data privacy concern regarding the dataset.

A.4 Availability of Dataset

Upon peer review of the work, we plan to publish the dataset. The dataset will be made publicly available to the research community upon the approval of the institutions involved.

A.5 Possible uses cases/improvements

One open question would be to keep looking for better models to model human behavior in facing moral dilemma. For such problems, one challenge is to get an interpretable model so that it has public acceptance which a black box solution is unlikely to have. Another interesting direction is to consider how to embed external ethical/philosophical constraints into a model and then learn constrained preferences according to gathered data. For example, a naive approach in constraining lexicographic preferences may be to fix priorities or preferences between some features beforehand based on constraints. Then the learning mechanism will only learn the preference relations between other features.

At the same time, we understand that as possible domains of moral dilemmas are limitless, it will be impractical to start gathering new data for every new moral dilemma. This motivates us to look into similarities of preferences under different moral dilemma. For example, we can gather similar data for multiple related scenarios and explore the possibility of transfer of knowledge between different models learnt from different data.