Exploring the Hierarhical Moral Principles Model

Sam Acquaviva (undergraduate, 9.66)

Abstract

Can a machine think morally? If so, can we align its moral values with our own? In this project, we explore a computational approach to these questions. We re-implement an existing model that predicts the moral values of a group and the moral values of individuals within that group. We provide new analysis into the model's inferred moral values. Then, we build and test a lower-dimensional representation of the model.

1 Introduction

Utilitarianism fundamentally relies on the belief that making a moral decision can be described by choosing the action that has the maximum utility [2]. In other words, people place weights (called *moral principles*) on different abstract features (e.g. age, gender). Then, when deciding which action is most moral, people choose the action which maximizes the trade-off between their moral principles.

Using this assumption that people base their moral decisions on the utility of the different outcomes, we can predict peoples moral principles by observing their decisions in different moral dilemmas. For example, if someone is repeatedly given the scenario where they have to choose between saving a dog and saving a cat, and they always choose to save the dog, then we can infer that their moral principles weight dogs more positively than cats.

Similarly, by looking at data from many individuals in the same group, we can infer the moral principles of that group. Building a model that can predict the moral values of a group and individuals within that group is interesting for a few reasons:

1. Effectively fitting people's moral choices will give us quantitative insight the moral values of a community, and how much individual moral values stray from that norm.
2. Designing intelligent machines that act ethically is an incredibly important task. Designing a flexible system that can model human moral decisions is a step in that direction.

2 Background

2.1 The Moral Machine

[1] builds an interface that presents users with moral dilemmas where they have to choose the lesser of two evils. In each scenario, a car has two choices: it can stay on course and hit either a group of people or a barrier (depending on the scenario), or it can swerve and hit a different group of people or a barrier (depending on the scenario). An example situation is shown in Figure 1.

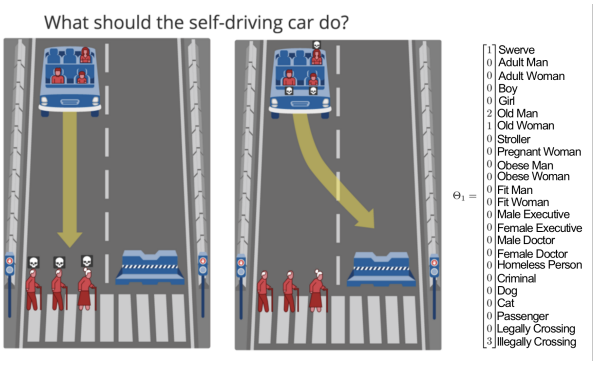

Figure 1: An example of a moral dilemma from the Moral Machine interface. On the right is the utility gained from choosing to swerve, Θ_1.

There are 20 different characters in the Moral Machine: an adult man, an adult woman, a boy, a girl, an old man, an old woman, a baby in a stroller, a pregnant woman, an obese man, an obese woman, a male athlete, a female athlete, a male businessman, a female businesswoman, a male doctor, a female doctor, a homeless person, a criminal, a dog, and a cat. In each scenario, there is a different number of each character type in either path, or one of the paths may have a barrier instead. If a barrier is in the path of the car, the passengers will die. If a group of people are in the path of the car, that group of people will die. Additionally, there may be a stop sign or walk sign that indicates whether the characters on the cross walk are legally or illegally crossing the street.

The utility of swerving can be represented by a K-dimensional vector Θ_1. In our representation, we will set $K = 24$. The first index will repre-

sent that we are swerving, the next 20 indices will represent the counts of each character type that is saved. The final 3 indices represent the number of people saved who are passengers, legally crossing, and illegally crossing, respectively. See Figure 1 for an example of Θ_1. we can construct the utility of continuing straight Θ_0 in a similar manner. We will use Θ to refer to (Θ_0, Θ_1)

2.2 The Hierarchical Bayesian Approach

[4] uses the data from the Moral Machine to build the *hierarchical moral principles* model (HMP), a hierarchical Bayesian model that predicts an individual's values and the value's of their community.

Instead of representing the utility of each scene as Θ, HMP maps each scene by a linear transformation A into a new D-dimensional vector $\Lambda = A\Theta$ that represents more abstract features of the scene ($D < K$).

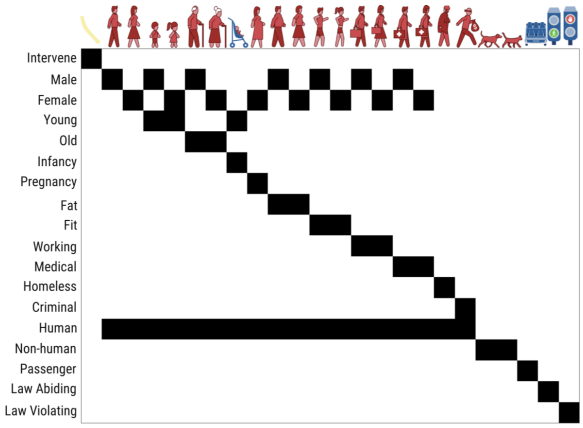

Figure 2: The A matrix used to map Θ to the more abstract representation Λ.

Using this representation of each scene, the hierarchical moral principles model builds a D-dimensional vector w_i that represents how much each dimension of Θ is valued by individual i. With w_i, we can represent the utility of a scenario Θ_i, according to the values in w_i, by calculating $u(\Theta_i) = w_i^T(A\Theta_i) = w_i^T\Lambda$. HMP models the probability of choosing to swerve as $P(Y = 1|\Theta) = F(u(\Theta_1) - u(\Theta_0))$, where $F(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

HMP importantly takes into account that an individual's morals are not independent from the morals of the group they are in. Each individual's moral principles w_i is sampled from a multivariate normal distribution with mean w^g and co-variance Σ^g, where w^g is the group's inferred moral principles and Σ^g is the variance within the group along each of the D abstract dimensions.

The group variance Σ^g is sampled from the LKJ co-variance matrix, and the group mean u^g is sampled from a 0-mean multivariate normal distribution with co-variance Σ^G. This entire graphical model is shown in Figure 3.

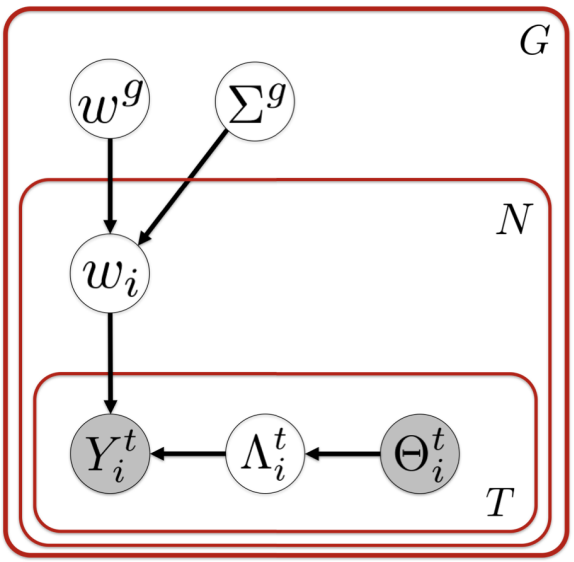

Figure 3: The full graphical model of HMP in plate notation.

3 Methods

3.1 Re-implementation

We re-implement the full hierarchical moral principles model, as well as three baselines proposed in [4]. For each baseline, and for the full model, we create a probabilistic program using the PyMC (Python Monte Carlo) library [6]. We use the same dataset as [4], which contains 130,000 decisions from 10,000 users. For each benchmark, we vary the training set size from $N = 4$ to $N = 128$ different users. For each user, we use their first 8 decisions to train, and their final 5 decisions to test the prediction accuracy.

None of the three benchmarks are hierarchical, and they all ignore the covariance between features. The first benchmark models the group weights w^g as sampled from a K-dimensional normal distribution with mean 0 and covariance $I\sigma^2$. So, this benchmark operates on the base parsing of the scene Θ and ignores the abstract dimensions Λ.

Benchmark 2 is the same as benchmark 1, but it operates on Λ instead of Θ. So, w^g is 18-dimensional.

Benchmark 3, instead of modeling only the average weights of the group, models only the weights

of each individual and assumes they are independent. Each w_i is modeled as w^g is modeled in benchmark 2 (drawn from a D-dimensional normal distribution with mean 0 and covariance $I\sigma^2$).

We then compare our results for these three benchmarks to the results from [4].

For a subset of the data collected from Denmark ($N = 99$), [4] also provides analysis on the inferred group covariance matrix which represents which abstract features are correlated for that group. We replicate this work and provide analysis of the covariance between the weights.

3.2 New Abstract Concepts

The linear-transformation matrix A that maps Θ to Λ is not the only possible abstraction. We propose a new transformation, A', that maps Θ to Λ' as shown in Figure 4. A' only models the five most important features in A, as determined by magnitude of that feature's value in HMP (Human, Young, Intervene, Old, Female).

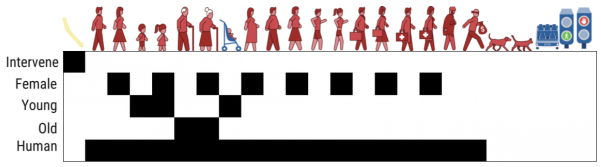

Figure 4: The A' matrix used to map Θ to the more abstract representation Λ'.

We compare the performance of HMP when Λ' to represent the utility of each scenario, and compare it against the performance of the benchmarks and of HMP when using Λ.

4 Results

4.1 Replication

4.1.1 Denmark Analysis

Our inferred group moral principles covariance matrix versus the group moral principles covariance matrix from [4] is shown in Figure 5. The weights seem qualitatively identical (we did not have access to the quantitative results from the original analysis).

The feature co-variances are very intuitive. For example, the most anti-correlated features are young versus old and fit versus large. Since these are antonyms, it is logical that when someone values one feature (i.e. fit) more heavily than average, they will value the other feature (i.e. large) equally in the other direction. On the other side, the most

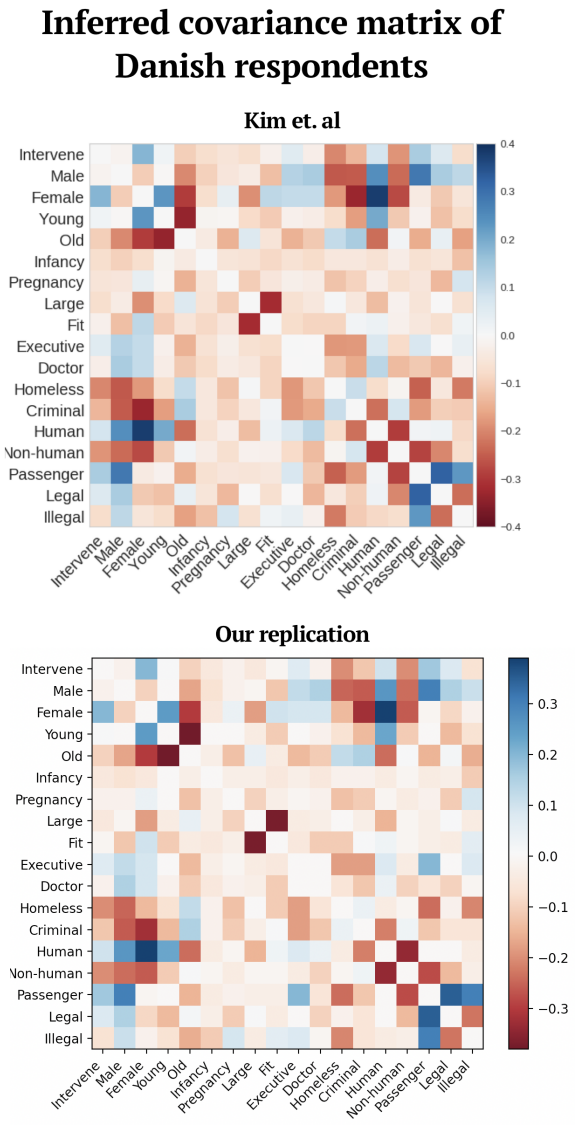

Figure 5: The inferred covariance matrices from the original hierarchical moral principles paper (left) and our replication of the model, on the same dataset.

correlated features are female and human. This is intuitive because all of the human characters are female. However, it is interesting to note that the covariance between human and female is larger than the covariance between human and male.

4.1.2 Benchmarks, Full Model, and Alternative A'

We graph the performance of the each benchmark from the original paper (for which we did have quantitative results) and from our replication in 6. We also include our results from the model using the abstract features Λ' instead of Λ. The given value at each N is the average over 30 random samples.

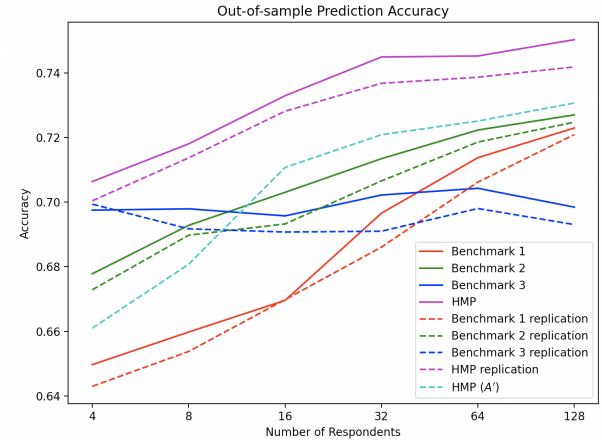

Figure 6: The out of sample accuracy of the benchmarks and full Hierarchical Moral Principles (HMP) model from Kim et. al (solid line), and the out of sample accuracy of our implementation of the benchmarks, full HMP model, and full HMP model with A' (dashed line).

5 Discussion

5.1 Performance of replication and A'

For all 3 benchmarks and for the full HMP model, our replication had slightly lower accuracy than the original implementation.

We offer a few possibilities. First, our parsing of each scene could have been slightly incorrect. The data from the original paper is not completely self-explanatory, so parsing the 130,000 decisions into the correct Θ required making educated inferences based on patterns in the data. For example, it was unclear whether 1 in the data represented a red light or green light. Since more people chose to save walkers when the light was in state 1 rather than state 2, we inferred that 1 indicates a green light and 2 represents a red light.

The difference could also be due to unspecified priors. More specifically, creating an LKJ covariance matrix in PyMC requires providing a distribution for the standard deviations. We used a half-normal distribution with $\mu = 0$ and $\sigma = 1$, but this could be different than the distribution used in the original paper.

HMP using Λ' instead of Λ had consistently worse accuracy. This is unsurprising, as Λ' is 5-dimensional while Λ is 18-dimensional, and Λ' is a subset of Λ, so it provides no new features. However, it did perform fairly well, outperforming all 3 benchmarks with $N = 128$. In future work, it would be interesting to see a more detailed ablation study of A, where we analyze the model performance when it doesn't have any subset of features.

5.2 What Didn't Work

In this project, the most difficult part was getting associated with using the probabilistic programming library PyMC. PyMC is built on top of Theano, an automatic differentiation framework that is no longer developed (although PyMC maintains it). Likely due to this, debugging is much more difficult than it would be if it was built on top of a better maintained framework. If we were to restart the project, we would instead use Pyro [3], which is built on top of PyTorch and seems to be much more popular.

Another roadblock was the sampling speed. Using a hierarchical model with increasing dimensionality at each layer seemed to make sampling very slow. Switching from a Monte Carlo sampler to the variational inference method ADVI (Automatic Differentiation Variational Inference) sped this up dramatically [5].

5.3 Risks of Model

In 1, we listed two motivating factors for this project: quantitative insights into people's moral choices and helping to create machines that act morally. However, it is important to note the difficulty with using HMP to help a machine make a moral choice.

Although the prediction accuracy of HMP shows that it can represent the moral principles of individuals and groups fairly well, it is a large assumption to assume that the action that maximizes utility is the most moral. Although we will not discuss this point further, it is important to note that philosophers have not unanimously decided that Utilitarianism is correct [7].

The moral machine is a very standardized environment, and using this technique in a real-world setting may not scale nicely. When noise is inevitably introduced (e.g. a sensor mistakenly senses 2 people instead of 3), are we still comfortable with the predictions that the model makes? Again, this is a question we ask without an answer.

5.4 Future Exploration

There are many possible avenues to extend this work. As mentioned in 5.1, one could continue exploration of A and analyze the performance of HMP when ablating features of A.

HMP shows how important modeling the hierarchy is in a Bayesian setting. However, it is not clear that this hierarchy needs to be explicitly modeled, as it is in HMP. In a neural network trained to

predict the moral principles of an individual, would it be possible to encourage the network to learn a hierarchical structure? If so, how could you analyze the hierarchical components after training?

Finally, another interesting area for research would be replacing HMP with a human to predict other people's decisions on the Moral Machine. Not only would this serve as a performance baseline for a computational model like HMP, but it could provide insight into a). how well people can infer the moral principles of others, and b). how human-like the error's of HMP are.

References

[1] 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, Nov 2018.

[2] Jeremy Bentham. An introduction to the principles of morals and legislation. *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation*, 1789.

[3] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[4] Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Josh Tenenbaum, and Iyad Rahwan. A computational model of commonsense moral decision making. *CoRR*, abs/1801.04346, 2018.

[5] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference, 2016.

[6] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016.

[7] JJC Smart and Bernard Utilitarianism Williams. *For and against*. Cambridge University Press, Cambridge, 1973.