A Computational Model of Commonsense Moral Decision Making

Richard Kim
Massachusetts Institute of Technology
Cambridge, Massachusetts
kimr@mit.edu

Max Kleiman-Weiner
Massachusetts Institute of Technology
Cambridge, Massachusetts
maxkw@mit.edu

Andrés Abeliuk
Massachusetts Institute of Technology
Cambridge, Massachusetts
aabeliuk@mit.edu

Edmond Awad
Massachusetts Institute of Technology
Cambridge, Massachusetts
awad@mit.edu

Sohan Dsouza
Massachusetts Institute of Technology
Cambridge, Massachusetts
dsouza@mit.edu

Joshua B. Tenenbaum
Massachusetts Institute of Technology
Cambridge, Massachusetts
jbt@mit.edu

Iyad Rahwan
Massachusetts Institute of Technology
Cambridge, Massachusetts
irahwan@mit.edu

ABSTRACT

We introduce a computational model for building moral autonomous vehicles by learning and generalizing from human moral judgments. We draw on a cognitively inspired model of how people and young children learn moral theories from sparse and noisy data and integrate observations made from different people in different groups. The problem of moral learning for autonomous vehicles is cast as learning how to weigh the different features of the dilemma using utility calculus, with the goal of making these trade-offs reflect how people make them in a wide variety of moral dilemma. By modeling the structures of individuals and groups in a hierarchical Bayesian model, we show that an individual's moral values – as well as a group's shared values – can be inferred from sparse and noisy data. We evaluate our approach with data from the Moral Machine, a web application that collects human judgments on moral dilemmas involving autonomous vehicles, and show that the model rapidly and accurately infers people's preferences and can predict the difficulty of moral dilemmas from limited data.

KEYWORDS

Artificial Intelligence; Machine Ethics; Moral Learning; Bayesian Inference

ACM Reference Format:
Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B. Tenenbaum, and Iyad Rahwan. 2018. A Computational Model of Commonsense Moral Decision Making. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3278721.3278770

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02…$15.00
https://doi.org/10.1145/3278721.3278770

1 INTRODUCTION

Recent advances in machine learning, notably deep learning, have demonstrated impressive results in various domains of human intelligence, such as computer vision [26], machine translation [30], and speech generation [21]. In domains as abstract as human emotion, deep learning has shown a proficient capacity to detect human emotions in natural language text [7]. These achievements might suggest that deep learning will also pave the way for AI in ethical decision making.

However, training deep learning models often requires large quantities of human-labeled data. Despite recent advances that enable models to be trained from a smaller number of examples [24, 29], this constraint remains a key challenge for deep learning. In addition, deep learning models have been criticized as "blackbox" algorithms that defy attempts at interpretation [16]. The viability of many deep learning algorithms for real-world applications in business and government has come into question as a recent legislation in the EU, slated to take effect in 2018, will ban automated decisions, including those derived from machine learning if they cause an "adverse legal effect" on the persons concerned [9].

In contrast to deep learning algorithms, evidence from studies in human cognition suggests that humans are able to learn and make predictions from a much smaller number of noisy and sparse examples [27]. Moreover, in the moral domain, people often make moral judgments for reasons they are able to articulate and explain the abstract principles that underlie their decision. Given this stark difference between the current state of machine learning and human cognition, how can we draw on the latest frameworks in cognitive science to design AI with the capacity to learn moral values from limited interactions with humans and make decisions with explicable processes?

A recent framework from the field of cognitive science postulates that humans learn to make moral decisions by acquiring values

along abstract moral concepts through observation and interaction with other humans in their environment [14]. This approach characterizes ethical decisions as the utility maximizing choice over a set of outcomes whose values are computed from weights people place on abstract moral concepts such as "kin" or "reciprocal relationship." In addition, given the dynamics of individuals and their memberships in a group, the framework depicts the process of how an individual's moral preferences, and the actions resulting from them, lead to a development of the group's shared moral principles (i.e. group norms).

In this work we extend the framework introduced by [14] to explore a computational model of learning preferences and human biases in moral decisions involving autonomous vehicles. We characterize moral judgment as a net utility maximizing decision over a function that computes trade-offs of values perceived by humans in the choices of the dilemma. These values are the weights that people place on abstract dimensions of the dilemma and we call these weights *moral principles*. Furthermore, we represent an individual agent as a member of a group with many other agents that are assumed to share similar moral principles; these shared moral principles of the group as an aggregate give rise to the *group norm*. Exploiting the hierarchical structure of individuals and group, we show how hierarchical Bayesian inference [8] can provide a powerful mechanism to rapidly infer individual preferences in moral decisions as well as the group norm from sparse and noisy data.

We apply our model to the domain of autonomous vehicles (AV) through a data set from the Moral Machine, a web application that collects human judgments in ethical dilemmas involving AV.[1] A recent study on public sentiment on AV reveals that endowing AI with human moral values is an important step before AV can undergo widespread market adoption [4]. In light of this study, we view application of our model to demonstrate inference of moral preferences in ethical decisions on the road as an important step towards building an AV with moral values acceptable to people.

This paper makes the following distinct contributions towards building an ethical AI for AVs:

- We explore a computational model of moral learning and show that inference of the parameter values over abstract features of moral dilemma enables faster learning of preferences and biases.
- Exploiting the social structure of individuals and groups as a hierarchical Bayesian model, we show that the inference over moral preferences of individuals, as well as those of the group, can be achieved rapidly from limited observations.
- Using response time as a proxy measure for difficulty of assessing a dilemma, we show that moral dilemmas have varying degrees of cognitive cost for the human judges, paving the way to incorporate confidence level of human judgment into inferring human preferences.

2 MORAL MACHINE DATA

Moral Machine is a web application built to collect and analyze human perceptions of moral dilemmas involving autonomous vehicles. As of October 2017, the application has collected over 30 million responses from over 3 million unique respondents from over 180 countries around the world. Here, we briefly describe the design of moral dilemma and data structure in Moral Machine.

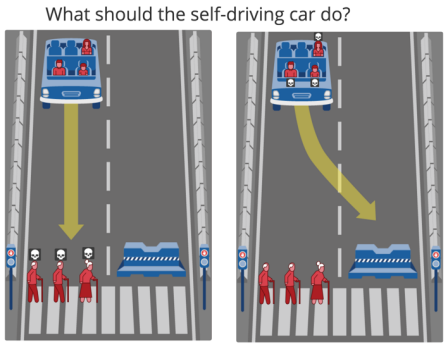

Figure 1: *Moral Machine* interface. An example of a moral dilemma that features an AV with sudden brake failure, facing a choice between either not changing course, resulting in the death of three elderly pedestrians crossing on a "do not cross" signal, or deliberately swerving, resulting in the death of three passengers; a child and two adults.

In a typical Moral Machine session, respondent is shown thirteen scenarios such as the example shown in Figure 1. In each scenario, the respondent is asked to choose one of two outcomes that have different ethical consequences with different trade-offs. A scenario can contain any random combination of twenty characters (see Figure 2) that represents various demographic attributes found in a general population. In addition to the demographic factors, a Moral Machine scenario also includes the factors of character's status as a passenger or a pedestrian and its status as a pedestrian who is crossing on green light or red light.

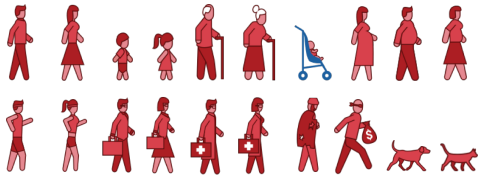

Figure 2: Twenty characters in Moral Machine represent various demographic attributes such as gender, age, social status, fitness level, and species.

In addition to the respondents' decisions, data about their response duration (in seconds) to each scenario and their approximate geo-location is also collected. This allows us to infer the country or region of access.

Every scenario has two choices, which we represent as a random variable Y with two realizable values $\{0, 1\}$. A respondent's choice to swerve (i.e., intervene) is represented as $Y = 1$, and likewise, their choice to stay (i.e., not-intervene) is represented as $Y = 0$. The respondent's choice yields a state in which certain set of characters are saved over the other. The resultant state is represented by character vector $\Theta_y \in \mathbb{N}^K$, which denotes the resultant state of choice y.

[1] http://moralmachine.mit.edu/

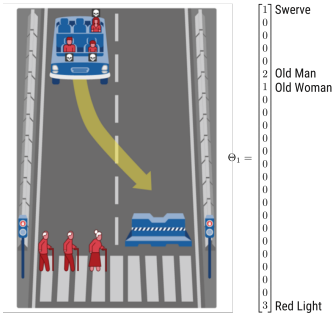

Figure 3: An example of vector representation of a state in the Moral Machine character space.

As an illustration, we show a vector representation of a resultant state of swerve in Figure 3. The vector element of *old man* character is denoted by value of 2, representing two *old man* characters that will be saved from the choice of swerve. In addition, the vector element of *red light* feature is denoted by value of 3, representing three pedestrians who are crossing the red light.

3 MORAL DILEMMA AS UTILITY FUNCTION

Jeremy Bentham, the founder of modern utilitarian ethics, described ethical decision in a moral dilemma as a utility maximizing decision over the sum of trade-offs over values in the dilemma [2]. More recently, cognitive psychologists have formalized the idea of analyzing moral dilemma using utility function that computes various trade-offs in the dilemma [18, 19]. Evidence of moral decision making in young children suggests that children base their moral judgments by computing trade-off of values over abstract concepts [15].

Here our aim is to model how a respondent arrives to his/her decision based on the values that he/she places on abstract dimensions of the moral dilemma, which we label *moral principles*. For instance, when a respondent chooses to save a female doctor character in a scenario over an adult male character, this decision is in part due to the value that respondent places on the abstract concept of *doctor*, a rare and valuable member in society who contributes to improvement of social welfare. The abstract concept of *female* gender also would be a factor in his or her decision.

In Moral Machine, twenty characters share many abstract features such as *female*, *elderly*, *non-human*, etc. Hence, the original character vector Θ_y can be decomposed into a new vector in the abstract feature space $\Lambda_y \in \mathbb{N}^D$ where $D \leq K$ via feature mapping $F : \Theta \rightarrow \Lambda$. In this work, we use a linear mapping $F(\Theta) = A\Theta$ where A is a 18×24 binary matrix such as the one shown in Figure 4.

Shown in Figure 5, the original state vector in the Moral Machine character space Θ is mapped on to a new state vector in the abstract feature space Λ. We note that vector element of *old* is denoted by value of 3 representing three character with this feature.

We define moral principles as weights $w \in \mathbb{R}^D$ that respondent place along the D abstract dimensions Λ. These weights represent how the respondent values abstract features such as *young*, *old*, or *doctor* to compute utility value of their choices. For simplicity, we

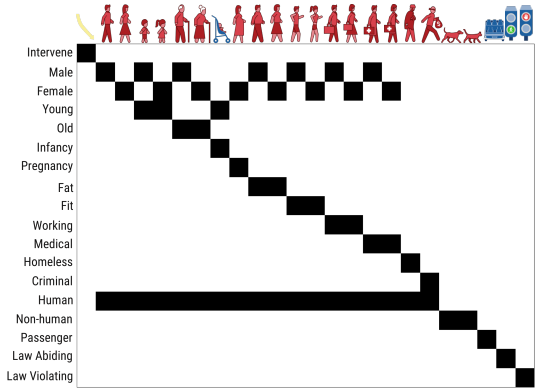

Figure 4: An example of a binary matrix A that decomposes the characters in Moral Machine into abstract features. Black squares indicate the presence of abstract features in the characters.

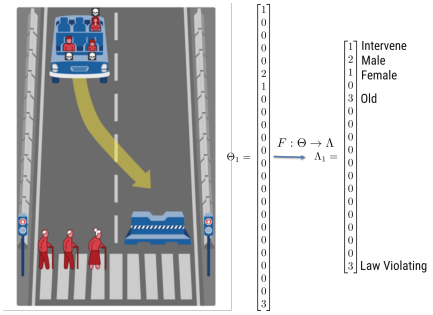

Figure 5: Vector representation of abstract features of a scenario choice.

model the utility value of a state as a linear combination of the features in the abstract dimension:

$$u(\Theta_i) = w^\top F(\Theta_i) \tag{1}$$

With utility values of the choice to not-intervene and intervene, respondent's decision to intervene $Y = 1$ is seen as probabilistic outcome based on sigmoid function of net utility of the two choices:

$$P(Y = 1|\Theta) = \frac{1}{1 + e^{-U(\Theta)}} \tag{2}$$

where

$$U(\Theta) = u(\Theta_1) - u(\Theta_0). \tag{3}$$

We turn our attention to inferring individual moral principles of respondents from sparse and noisy observation of their decisions in moral dilemma.

4 HIERARCHICAL MORAL PRINCIPLES

Studies by anthropologists have shown that societies across different regions and time periods hold widely divergent views about what actions are ethical [3, 12, 13]. For example, certain societies strongly emphasize respect for the elderly while others focus on protecting the young. These views in a society are what we refer to as the society's *group norms*.

Nevertheless, even in a society with a homogeneous cultural and ethnic make-up, individual members of the group can hold unique and different moral standards [11]. How can we model the complex relationship between the group norm and individual moral principles?

We introduce *hierarchical moral principles* model, which is an instance of hierarchical Bayesian model [8]. Returning to data in Moral Machine, consider N respondents that belong to a group $g \in G$. This group can be a country, a culture, or a region within which customs and norms are shared.

The moral principles of respondent i is drawn from a multivariate normal distribution parameterized by the mean values of the group w^g on the D dimensions:

$$w_i \sim Normal_D(w^g, \Sigma^g), \tag{4}$$

where the diagonal of the covariance matrix Σ^g represents the in-group variance or differences between the members of the group along the abstract dimensions. Higher variance value describes broader diversity of opinions along that corresponding abstract dimension. In addition, covariance (off-diagonal) value captures the strength of relationship between the values they place on abstraction dimension. As an example, a culture that highly values *infancy* should also highly value *pregnancy* as they are intuitively closely related concepts. Covariance matrix allows the Bayesian learner to understand related concepts and use the relationship to rapidly approximate the values of one dimension after inferring that of a highly correlated dimension.

Let $\mathbf{w} = \{w_1, ..., w_i, ..., w_N\}$ be a set of unique moral principles by N respondents. Each respondent i makes judgments on T scenarios $\mathbf{\Theta} = \{\Theta_1^1, ..., \Theta_i^t, ..., \Theta_N^T\}$. Judgment by respondent i is an instance of a random variable Y_i^t. Given the observation of the set of states $\mathbf{\Theta}$ and the decisions $\mathbf{Y}$, the posterior distribution over the set of moral principles follows:

$$P(\mathbf{w}, w^g, \Sigma^g | \mathbf{\Theta}, \mathbf{Y}) \propto P(\mathbf{\Theta}, \mathbf{Y} | \mathbf{w}) P(\mathbf{w} | w^g, \Sigma^g) \\ P(w^g) P(\Sigma^g) \tag{5}$$

where the likelihood is

$$P(\mathbf{\Theta}, \mathbf{Y} | \mathbf{w}) = \prod_{i=1}^{N} \prod_{t=1}^{T} p_{ti}^{y_i^t} (1 - p_{ti})^{(1-y_i^t)} \tag{6}$$

and $p_{ti} = P(Y_i^t = 1 | \Theta^t)$ is the probability that a respondent chooses to swerve in scenario t given Θ^t as shown in Equation 2. Graphical representation of the model is presented in Figure 6.

As an illustration, we randomly sampled 99 respondents from Denmark, which equates to 1,287 response data. We specified prior over the covariance matrix $P(\Sigma^g)$ with LKJ covariance matrix [17] with parameter $\eta = 2$:

$$\Sigma^g \sim LKJ(\eta) \tag{7}$$

and the prior over group weights $P(w^g)$ with

$$w^g \sim Normal_D(\mu, \Sigma^g) \tag{8}$$

where $\mu = \mathbf{0}$.

We inferred the individual moral principles as well as the group values w^g and the covariance matrices Σ^g. These results are shown in Figure 7. We note the variations in the inferred moral principles of three representative sub-sample of Danish respondents.

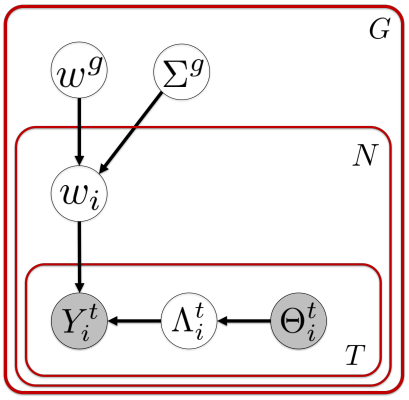

Figure 6: Graphical representation of hierarchical Bayesian model of moral principles.

4.1 Predicting Individual Judgments

As a means to evaluate our model, we performed out-of-sample prediction test. We randomly selected ten-thousand respondents from the Moral Machine website who completed at least one session, which contains thirteen scenarios. We filtered only the respondents' first thirteen scenarios to compile a data set consisting 130,000 decisions.

We compared the predictive accuracy of the model against three benchmarks. Benchmark 1 models the collective values of the characters in Moral Machine such that the utility of a state is computed as

$$u(\Theta) = w^{c\top} \Theta \tag{9}$$

where $w^c \in \mathbb{R}^K$. Benchmark 1 models the weights as

$$w^c \sim Normal_K(\mu, \sigma^2 I) \tag{10}$$

and does not include the group hierarchy or the covariance between the weights over the characters and factors (e.g. traffic light, passenger, etc.).

Benchmark 2, which builds upon Benchmark 1, models the values along the abstracts moral dimensions Λ as $w^f \sim Normal_D(\mu, \sigma^2 I)$. The group hierarchy and the covariance between weights are ignored.

Finally, benchmark 3 models the individual moral principles of each respondent as $w_i^l \sim Normal_D(\mu, \sigma^2 I)$, but does not include the hierarchical structure. Therefore, each respondent is viewed as an independent agent wherein inferring the values of one respondent provides no insight about the values of another.

To demonstrate the gains in accuracy, we tested the models across different size of training data by varying the number of sampled respondents along $N = (4, 8, 16, 32, 64, 128)$. We used the first eight judgments from each respondent as training data, and tested the accuracy of predictions on the remaining five of the responses per each agent. For our model, we assumed that sampled respondents of size N belong to one group.

The results (Figure 8) shows that as the number of respondents (i.e. training data) grows larger, predictive accuracy of our model, benchmark 1 and 2 improve. Accuracy of benchmark 3 does not improve as the the number of respondents have no bearing on

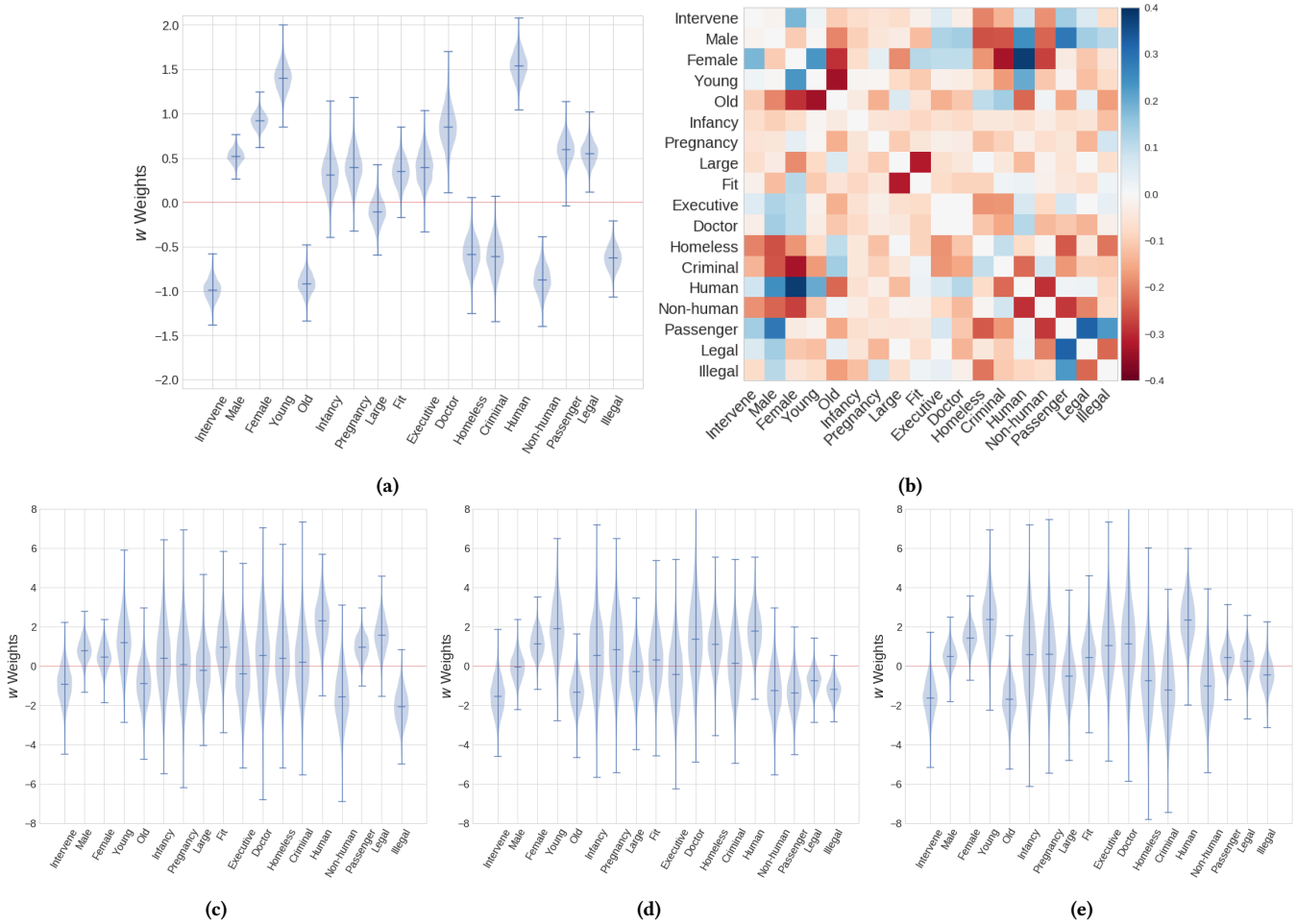

Figure 7: (a) Inferred group norm of sampled Danish respondents; (b) Inferred covariance matrix of the Danish respondents; (c-e) Individual moral principle values of three representative sub-sample of Danish respondents.

inference of individual respondent's values. However, the hierarchical moral principles model shows consistently improving accuracy rates along the increasing size of the training data.

We note that the margin of improvement between benchmark 1 and benchmark 2 reveals the gain achieved from abstraction and dimension reduction. The margin between benchmark 2 and our model reveals the gain from including individual moral principles. Finally, the margin between benchmark 3 and our model is indicative of the gain achieved by the group hierarchy.

4.2 Response Time

Studies in human decision making find strong relationship between the confidence level of the decision and reaction time of the decision (i.e. reaction time) [1, 5, 25]. These studies show that human subjects in binary-decision tasks take longer time to arrive at a decision when there is lower level of evidence. In this section, we take this approach to show that our model accurately captures the relationship between reaction time and difficulty of a moral dilemma.

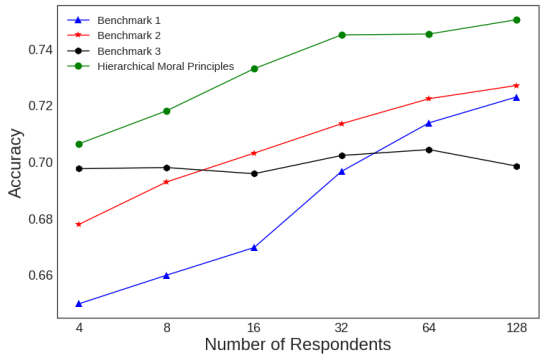

Figure 8: Comparison of out-of-sample prediction accuracy rates of the hierarchical moral principles model and the three benchmark models.

We sampled 1727 respondents who accessed Moral Machine from the US; which altogether correspond to 22,451 judgments. In

addition to the judgment decisions, we measured response times (RT) in seconds that the respondents took to arrive at their decisions. Due to the unsupervised nature of the experiment, respondents are free to stop and reengage at later time; as such, we eliminated responses that took more than 120 seconds from our analysis. From the judgment data, after inferring the moral principles of individual respondents, we computed the estimated the probability of decision to swerve (i.e. $p_i^t = P(Y_i^t = 1|\Theta_i^t)$) of each scenario as defined in Equation 2. We computed new metric, *certainty of decision*, using $|p_i^t - 0.5|$.

Plotting the certainty of decision and response times of the scenarios (see Figure 9) reveals an intuitive pattern of relationship between two variables.

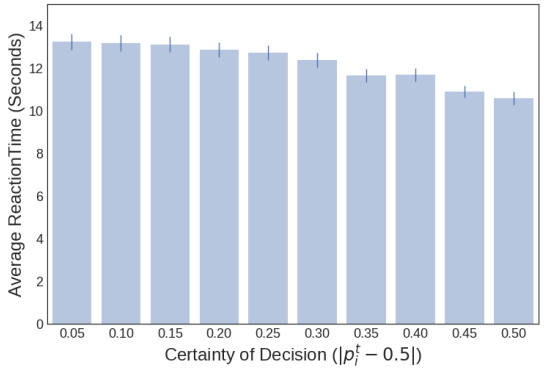

Figure 9: Reaction time in seconds per estimated certainty of decisions, which is defined as distance of the probability of judgment from the 0.5 probability of swerving.

Scenarios with higher certainty represent those that have clear trade-offs in the dilemmas such that the respondents on average respond quicker to the dilemmas. Likewise, scenarios with lower certainties are those that have ambiguous trade-offs such that the respondents have less confidences about their decisions. Intuitively, resolving the ambiguity of the trade-offs takes greater cognitive costs, which is revealed as longer response times for the respondents.

We view the relationship between response time and estimated certainty of decision from the model as a supporting evidence that the model is a robust representation of how people resolves moral dilemmas. In addition, the fact that cognitive cost of value-based decision process is revealed in their reaction times is an extra bit of information that could be used in the inference. For instance, we see a person making a quick decision; then we might also get information about the relative value difference between the two choices. In future work, we intend to integrate response time information into the process of learning itself to allow the learner to infer even faster.

5 DISCUSSION

Drawing on a recent framework for modeling moral learning, we proposed a computational model of inferring biases and preferences of human decision makers in moral dilemmas. We demonstrated the application of this model in the domain of autonomous vehicles using data from Moral Machine. We showed that hierarchical Bayesian inference provides a powerful mechanism to accurately infer individual preferences as well as group norms along abstract dimensions. We concluded with a demonstration of the model successfully capturing the cognitive cost in resolving the trade-offs in moral dilemmas. We show that moral dilemmas that are difficult to predict human judgments according to the model are correlated with long response times, in which response times serve as proxy for difficulty of the dilemma.

In this work, we have left out any normative discussion of how to aggregate the individual moral principles and the group norms to design an AI agent that makes decisions that optimize social utility of all other agents in the system. Recently [20] introduced a novel method of aggregating individuals preferences such that the decision reached after the aggregation ensures global utility maximization. We view this method as a natural complement to our work.

Another interesting extension of our work is to explore the mechanism that maps the observable data on to the abstract feature space. We formalized this process as feature mapping $F : \Theta \rightarrow \Lambda$. Evidence from developmental psychology suggests that children grow to acquire abstract knowledge and form inductive constraints [6, 10]. Non-parametric Bayesian processes such as the Indian Buffet Process [28] and its variants [22] are promising models to learn the feature mapping as well in the moral domain.

We used response time as a proxy to measure decision difficulty and proposed that the response time can be used as extra information for more accurate inference over respondent's individual moral principles. Combining our current model with drift diffusion model [23] can lead to a richer model that describes confidence and error in moral decision making. An AI agent needs to understand the moral basis of people's actions including when they are from socially inappropriate moral values as well as when they are simply mistakes made too quickly. For instance, if an AI agent observes a person who spends a long time to make a ultimately wrong decision, the AI agent should incorporate the person's confidence level and error rates to make accurate inference that the person may have been likely to have made a mistake.

Finally, we used the same source of data for both inference of the abstract moral principles and for testing the model's predictive power. However, the abstract dimensions of the characters and factors in Moral Machine are not confined to the Moral Machine data set nor even the AV domain. An interesting experiment would be to test the model across various moral dilemmas in different contexts. Hierarchical Bayesian models like one studied here have been used successfully in transfer learning. Demonstrating the capacity to learn moral principles from one domain and applying these principles to ethical decisions in other domains is a key challenge for the development of human-like ethical AI.

REFERENCES

[1] Jonathan Baron and Burcu Gürçay. 2017. A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition* 45, 4 (5 2017), 566–575. https://doi.org/10.3758/s13421-016-0686-8

[2] Jeremy Bentham. 1789. *An Introduction to the Principles of Morals and Legislation.* https://doi.org/10.1111/j.2048-416X.2000.tb00070.x

[3] P. R. Blake, K. McAuliffe, J. Corbit, T. C. Callaghan, O. Barry, A. Bowie, L. Kleutsch, K. L. Kramer, E. Ross, H. Vongsachang, R. Wrangham, and F. Warneken. 2015. The ontogeny of fairness in seven societies. *Nature* 528, 7581 (2015), 258–261. https://doi.org/10.1038/nature15703

[4] Jean-FranÃğois Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (6 2016), 1573 LP – 1576. http://science.sciencemag.org/content/352/6293/1573.abstract

[5] Nicholas Cain and Eric Shea-Brown. 2012. Computational models of decision making: Integration, stability, and noise. https://doi.org/10.1016/j.conb.2012.04.013

[6] Susan. Carey. 2009. *The origin of concepts*. Oxford University Press. 598 pages. https://global.oup.com/academic/product/the-origin-of-concepts-9780199838806#.WfDc448zKVM.mendeley

[7] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] A Gelman, J B Carlin, H S Stern, D B Dunson, A Vehtari, and D B Rubin. 2013. *Bayesian Data Analysis, Third Edition*. Taylor & Francis. https://books.google.com/books?id=ZXL6AQAAQBAJ

[9] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". (6 2016). http://arxiv.org/abs/1606.08813

[10] Alison Gopnik and Andrew N Meltzoff. 1997. *Words, thoughts, and theories*. The MIT Press, Cambridge, MA, US. 268, xvi, 268–xvi pages.

[11] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* (2009). https://doi.org/10.1037/a0015141

[12] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review* 91, 2 (2001), 73–78. http://www.jstor.org/stable/2677736

[13] Bailey R. House, Joan B. Silk, Joseph Henrich, H. Clark Barrett, Brooke A. Scelza, Adam H. Boyette, Barry S. Hewlett, Richard McElreath, and Stephen Laurence. 2013. Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences* 110, 36 (2013), 14586–14591. https://doi.org/10.1073/pnas.1221217110

[14] Max Kleiman-Weiner, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Learning a commonsense moral theory. *Cognition* 167 (2017), 107–123. https://doi.org/10.1016/j.cognition.2017.03.005

[15] L. Kohlberg. 1981. Essays in Moral Development. In *the Philosophy of Moral Development*.

[16] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. (6 2016). http://arxiv.org/abs/1606.04155

[17] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100, 9 (2009), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

[18] John Mikhail. 2007. Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences* 11, 4 (2007), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

[19] John Mikhail. 2011. *Elements of Moral Cognition*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511780578

[20] Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. 2017. A Voting-Based System for Ethical Decision Making. (9 2017). http://arxiv.org/abs/1709.06692

[21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. (2016), 1–15. http://arxiv.org/abs/1609.03499

[22] Piyush Rai and Hal Daumé. 2009. The Infinite Hierarchical Factor Regression Model. *Advances in Neural Information Processing Systems 21* (2009), 1321–1328. http://arxiv.org/abs/0908.0570

[23] Roger Ratcliff and Gail McKoon. 2008. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation* 20, 4 (4 2008), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

[24] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. One-shot Learning with Memory-Augmented Neural Networks. (5 2016). http://arxiv.org/abs/1605.06065

[25] Philip L Smith and Roger Ratcliff. 2004. Psychology and neurobiology of simple decisions. *Trends in neurosciences* (2004). https://doi.org/10.1016/j.tins.2004.01.006

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. [n. d.]. Going Deeper with Convolutions. ([n. d.]).

[27] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331, 6022 (3 2011), 1279. http://science.sciencemag.org/content/331/6022/1279.abstract

[28] Zoubin Ghahramani Thomas L. Griffiths. 2005. Infinite Latent Feature Models and the Indian Buffet Process. *Advances in Neural Information Processing Systems 18* (2005), 475âĂŞ482. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.3951

[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. (6 2016). http://arxiv.org/abs/1606.04080

[30] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ÅĄukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. (9 2016). http://arxiv.org/abs/1609.08144