A Computational Model of Commonsense Moral Decision Making

Richard Kim **Max Kleiman-Weiner** **Andrés Abeliuk**

Edmond Awad **Sohan Dsouza** **Josh Tenenbaum** **Iyad Rahwan**

Massachusetts Institute of Technology, Cambridge MA, USA

Abstract

We introduce a new computational model of moral decision making, drawing on a recent theory of commonsense moral learning via social dynamics. Our model describes moral dilemmas as utility function that computes trade-offs in values over abstract moral dimensions, which provide interpretable parameter values when implemented in machine-led ethical decision-making. Moreover, characterizing the social structures of individuals and groups as a hierarchical Bayesian model, we show that a useful description of an individual's moral values – as well as a group's shared values – can be inferred from a limited amount of observed data. Finally, we apply and evaluate our approach to data from the Moral Machine, a web application that collects human judgments on moral dilemmas involving autonomous vehicles.

Recent advances in machine learning, notably Deep Learning, have demonstrated impressive results in various domains of human intelligence, such as computer vision (Szegedy et al.), machine translation (Wu et al., 2016), and speech generation (Oord et al., 2016). In domains as abstract as human emotion, Deep Learning has shown a proficient capacity to detect human emotions in natural language text (Felbo et al., 2017). These achievements indicate that Deep Learning will be paving the way for AI in ethical decision making.

However, training Deep Learning models often requires human-labeled data numbering in the millions, and despite recent advances that enables a model to be trained from a small number of examples (Vinyals et al., 2016; Santoro et al., 2016), this constraint remains a key challenge in Deep Learning. In addition, Deep Learning models have been criticized as "blackbox" algorithms that defy attempts at interpretation (Lei, Barzilay, and Jaakkola, 2016). 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 (Goodman and Flaxman, 2016).

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 (Tenenbaum et al., 2011). Moreover, studies have shown that humans are able to internally rationalize their moral decisions and articulate reasons for these (Haidt, 2001). Given this stark difference between the current state of machine learning and human cognition, how can we draw on the latest theories 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 theory from the field of cognitive science postulates that humans learn to make ethical decisions by acquiring abstract moral principles through observation and interaction with other humans in their environment (Kleiman-Weiner, Saxe, and Tenenbaum, 2017). This theory characterizes ethical decision as 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 explains 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 Kleiman-Weiner, Saxe, and Tenenbaum (2017) to explore a computational model of the human mind in moral dilemmas with binary-decisions. We characterize the decision making in moral dilemmas as a utility function that computes the trade-offs of values perceived by humans in the choices of the dilemma. These values are the weights that humans put on abstract dimensions of the dilemma; we call these weights *moral principles*. Furthermore, we represent an individual agent as a member of a group with many other agents that 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 (Gelman et al., 2013) can provide a powerful mechanism to rapidly infer individual moral principles as well as the group norm with 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 dilem-

mas 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 (Bonnefon, Shariff, and Rahwan, 2016). In light of this study, we view application of our model to understand how the human mind perceives and resolves moral dilemmas on the road as an important step towards building an AV with human moral values.

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

- Introducing a novel computational model of moral decision making that characterizes moral dilemma as a trade-off of values along abstract moral dimensions. We show that this model well-describes how the human mind processes moral dilemmas and provides an interpretable process for an AI agent to arrive at a decision in a moral dilemma.
- Characterizing the social structure of individuals and groups as a hierarchical Bayesian model, we show that the model can rapidly infer moral principles of individuals from limited number of observational data. Rapidly inferring other agents' unique moral values will be crucial, as AI agents interact with other agents, including humans.
- Demonstrating the model's capacity to rapidly infer group's norms, characterized as prior over individual moral preferences. Inferring shared moral values of a group is an important step towards designing an AI agent that makes socially optimal choices.

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.

In a typical Moral Machine session, a respondent is shown 13 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, 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.

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 of certain set of characters being saved over others. 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/

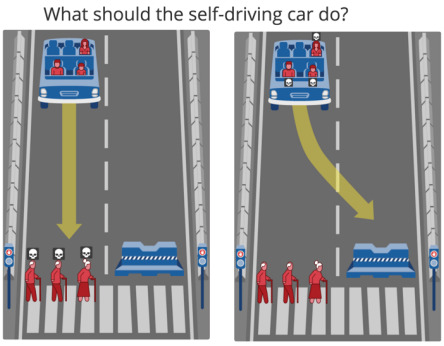

Figure 1: *Moral Machine* interface. An example of a moral dilemma that features an AV with sudden brake failure, facing a choice between either not changing course, resulting in the death of three elderly pedestrians crossing on a "do not cross" signal, or deliberately swerving, resulting in the death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

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

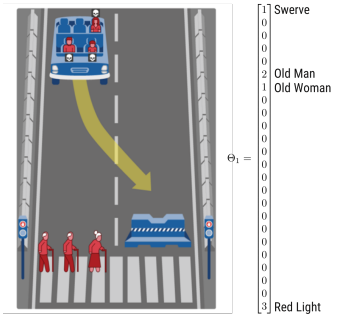

Figure 3: An example of vector representation of a state in the Moral Machine character space.

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 $Y = 1$. In addition, the vector element of *red light* feature is denoted by value of 3, representing three pedestrian who are

crossing the red light.

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 (Bentham, 1789). More recently, cognitive psychologists have formalized the idea of analyzing moral dilemma using utility function that computes various trade-offs in the dilemma (Mikhail, 2007, 2011). Evidence of moral decision making in young children suggests that children base their moral judgments by computing trade-off of values over abstract concepts (Kohlberg, 1981).

Using this framework, we can analyze 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.

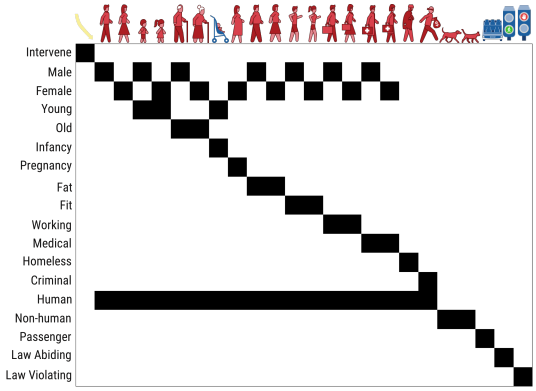

Figure 4: An example of a binary matrix A that decomposes the characters in Moral Machine into abstract features. Black squares indicate the presence of abstract features in the characters.

Shown in Figure 5, the original state vector in the Moral Machine character space Θ is mapped into 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 model the utility value of a state as a linear combination of the features in the abstract dimension:

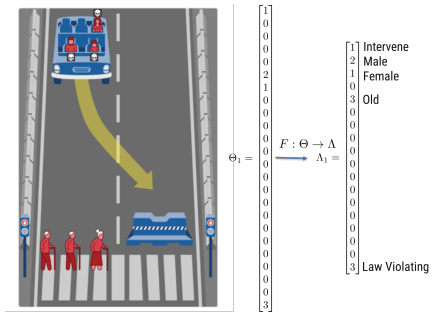

Figure 5: Vector representation of abstract features of a scenario choice.

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

With utility values of the choice to not-intervene $u(\Theta_0)$ and intervene $u(\Theta_1)$, 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.

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 (Henrich et al., 2001; House et al., 2013; Blake et al., 2015). 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 (Graham, Haidt, and Nosek, 2009). 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 (Gelman et al., 2013). 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) values capture 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.

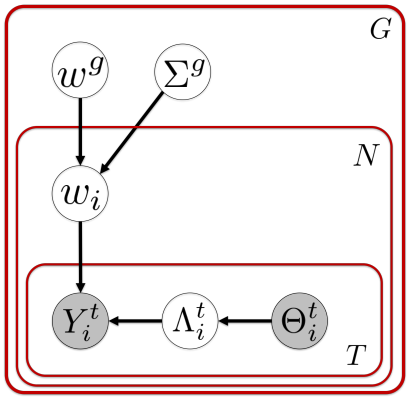

Figure 6: Graphical representation of hierarchical Bayesian model of moral principles.

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 (Lewandowski, Kurowicka, and Joe, 2009) 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.

Predicting Individual Judgments

As an evaluation of 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 13 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 $w^c \sim Normal_K(\mu, \sigma^2 I)$ 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 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.

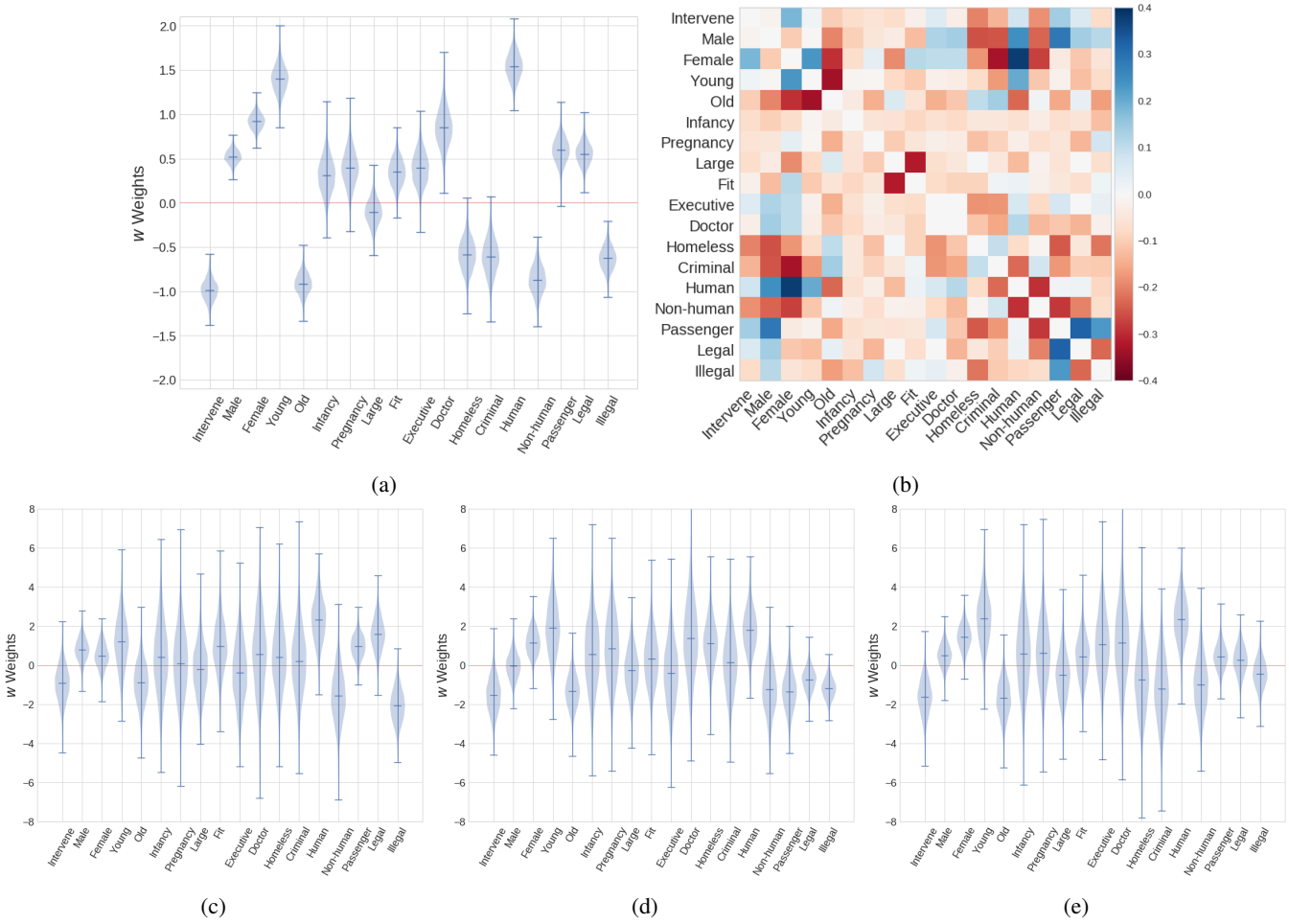

Figure 7: (a) Inferred group norm of sampled Danish respondents; (b) Inferred covariance matrix of the Danish respondents; (c-e) Individual moral principle values of three representative sub-sample of Danish respondents.

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) (Smith and Ratcliff, 2004; Cain and Shea-Brown, 2012; Baron and Gürçay, 2017). 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.

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 experiments, 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.

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 mak-

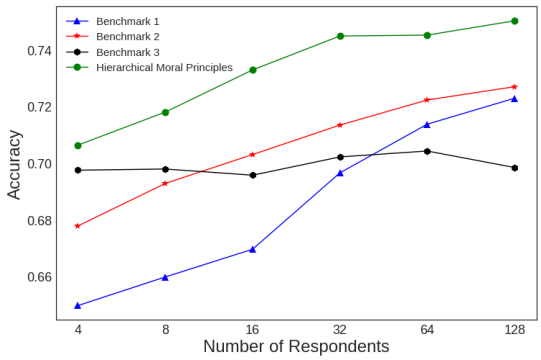

Figure 8: Comparison of out-of-sample prediction accuracy rates of the hierarchical moral principles model and the three benchmark models.

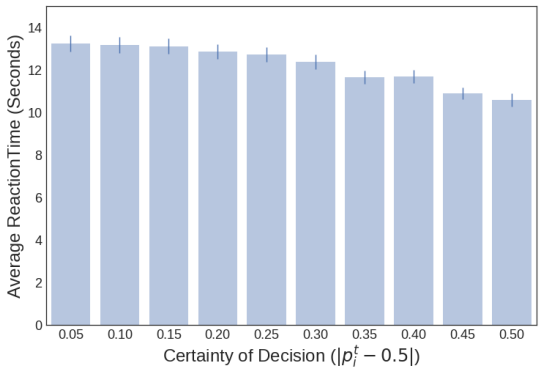

Figure 9: Reaction time in seconds per estimated certainty of decisions, which is defined as distance of the probability of judgment from the 0.5 probability of swerving.

ing 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.

Discussion

Drawing on a recent framework for modeling human moral decisions, we proposed a computational model of how the human mind arrives at a decision in moral dilemma. 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 the abstract moral 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 unpredictable by the model are correlated with long response times, where response times are a proxy of how difficult the dilemma is for the respondent because the subject is indifferent between the two responses.

In this work, we have left out any discussion about method 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, a paper by Noothigattu et al. (2017) 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 naturally 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 (Gopnik and Meltzoff, 1997; Carey, 2009). Non-parametric Bayesian processes such as the Indian Buffet Process (Thomas L. Griffiths, 2005) and its variants (Rai and Daumé, 2009) are promising models to characterize this learning mechanism in the moral domain.

We used response time as a proxy to measure cognitive cost and proposed that the response time can be used as an extra information for more accurate inference over respondent's individual moral principles. Combining our current model with drift diffusion model (Ratcliff and McKoon, 2008) can lead to a richer model that describes confidence and error in moral decision making. An AI agent needs to understand moral basis of people's actions including when they are from socially inappropriate moral values as well as when they are mistakes. 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 most likely made a mistake.

Finally, we have inferred abstract moral principles and tested the model's predictive power on the same source of data. However, the abstract dimensions of the characters and factors in Moral Machine are not confined to the AV domain. An interesting experiment would be to test the model across various moral dilemmas in difference contexts. Hierarchical Bayesian models has been applied in the domain of transfer learning. Demonstrating capacity to learn moral principles from one domain and apply these principles in other domains to make ethical decisions would show that a development of human-like ethical AI system does not need to be domain specific.

References

Baron, J., and Gürçay, B. 2017. A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition* 45(4):566–575.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*.

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

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

Cain, N., and Shea-Brown, E. 2012. Computational models of decision making: Integration, stability, and noise.

Carey, S. 2009. *The origin of concepts*. Oxford University Press.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 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)*.

Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; and Rubin, D. B. 2013. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Goodman, B., and Flaxman, S. 2016. European Union regulations on algorithmic decision-making and a "right to explanation".

Gopnik, A., and Meltzoff, A. N. 1997. *Words, thoughts, and theories.* Learning, development, and conceptual change. Cambridge, MA, US: The MIT Press.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*.

Henrich, J.; Boyd, R.; Bowles, S.; Camerer, C.; Fehr, E.; Gintis, H.; and McElreath, R. 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review* 91(2):73–78.

House, B. R.; Silk, J. B.; Henrich, J.; Barrett, H. C.; Scelza, B. A.; Boyette, A. H.; Hewlett, B. S.; McElreath, R.; and Laurence, S. 2013. Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences* 110(36):14586–14591.

Kleiman-Weiner, M.; Saxe, R.; and Tenenbaum, J. B. 2017. Learning a commonsense moral theory. *Cognition* 167:107–123.

Kohlberg, L. 1981. Essays in Moral Development. In *the Philosophy of Moral Development*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions.

Lewandowski, D.; Kurowicka, D.; and Joe, H. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100(9):1989–2001.

Mikhail, J. 2007. Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences* 11(4):143–152.

Mikhail, J. 2011. *Elements of Moral Cognition*. Cambridge: Cambridge University Press.

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2017. A Voting-Based System for Ethical Decision Making.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. 1–15.

Rai, P., and Daumé, H. 2009. The Infinite Hierarchical Factor Regression Model. *Advances in Neural Information Processing Systems 21* 1321–1328.

Ratcliff, R., and McKoon, G. 2008. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation* 20(4):873–922.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. One-shot Learning with Memory-Augmented Neural Networks.

Smith, P. L., and Ratcliff, R. 2004. Psychology and neurobiology of simple decisions. *Trends in neurosciences*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. Going Deeper with Convolutions.

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331(6022):1279.

Thomas L. Griffiths, Z. G. 2005. Infinite Latent Feature Models and the Indian Buffet Process. *Advances in Neural Information Processing Systems 18* 475482.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, .; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.