Using Machine Learning to Guide Cognitive Modeling: A Case Study in Moral Reasoning

Mayank Agrawal (mayank.agrawal@princeton.edu)
Department of Psychology, Princeton University

Joshua C. Peterson (peterson.c.joshua@gmail.com)
Department of Computer Science, Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University

Abstract

Large-scale behavioral datasets enable researchers to use complex machine learning algorithms to better predict human behavior, yet this increased predictive power does not always lead to a better understanding of the behavior in question. In this paper, we outline a data-driven, iterative procedure that allows cognitive scientists to use machine learning to generate models that are both interpretable and accurate. We demonstrate this method in the domain of moral decision-making, where standard experimental approaches often identify relevant principles that influence human judgments, but fail to generalize these findings to "real world" situations that place these principles in conflict. The recently released Moral Machine dataset allows us to build a powerful model that can predict the outcomes of these conflicts while remaining simple enough to explain the basis behind human decisions.

Keywords: machine learning; moral psychology

Introduction

Explanatory and predictive power are hallmarks of any useful scientific theory. However, in practice, psychology tends to focus more on explanation (Yarkoni & Westfall, 2017), whereas machine learning is almost exclusively aimed at prediction. The necessarily restrictive nature of laboratory experiments often leads psychologists to test competing hypotheses by running highly-controlled studies on tens or hundreds of subjects. Although this procedure gives a better understanding of the specific phenomenon, it can be difficult to generalize the findings and predict behavior in the "real world," where multiple factors are interacting with one another. Conversely, machine learning takes full advantage of complex, nonlinear models that excel in tasks ranging from image classification (Krizhevsky et al., 2012) to video game playing (Mnih et al., 2015). The performance of these models scales with their level of expressiveness (Huang et al., 2018), which results in millions of parameters that are difficult to interpret.

Interestingly, machine learning has long utilized insight from cognitive psychology and neuroscience (Rosenblatt, 1958; Sutton & Barto, 1981; Ackley et al., 1985; Elman, 1990), a trend that continues to this day (Banino et al., 2018; Lzaro-Gredilla et al., 2019). We believe that the reverse direction has been underutilized, but could be just as fruitful. In particular, psychology could leverage machine learning to improve both the predictive and explanatory power of cognitive models. We propose a method (summarized in Figure 1) that enables cognitive scientists to use large-scale behavioral datasets to construct interpretable models that rival the performance of complex, black-box algorithms.

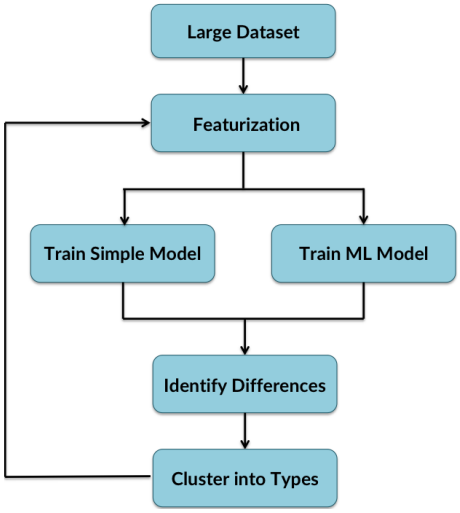


Figure 1: A systematic, data-driven procedure for building interpretable models that rival the predictive power of complex machine learning models.

This methodology is inspired by Box's loop (Box & Hunter, 1962; Blei, 2014; Linderman & Gershman, 2017), a systematic process of integrating the scientific method with exploratory data analysis. Our key insight is that training a black-box algorithm gives a sense of how much variance in a certain type of behavior can be predicted. This predictive power provides a standard for improvement in explicit cognitive models (Khajah et al., 2016). By continuously critiquing an interpretable cognitive model with respect to these black-box algorithms, we can identify and incorporate new features until its performance converges, thereby jointly maximizing our two objectives of explanatory and predictive power.

In this paper, we demonstrate this methodology by building a statistical model of moral decision-making. Philosophers and psychologists have historically conducted thought experiments and laboratory studies isolating individual principles responsible for human moral judgment (e.g. consequentialist ones such as harm aversion or deontological ones such as not using others as a means to an end). However, it can be difficult to predict the outcomes of situations in which these principles conflict (Cushman et al., 2010). The recently released

Moral Machine dataset (Awad et al., 2018) allows us to build a predictive model of how humans navigate these conflicts over a large problem space. We start with a basic rational choice model and iteratively add features until its accuracy rivals that of a neural network, resulting in a model that is both predictive and interpretable.

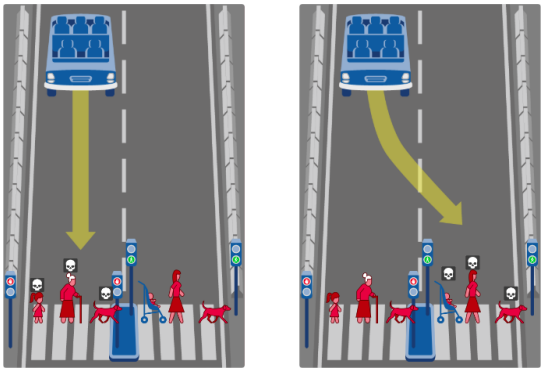
Background

Theories of Moral Decision-Making The two main families of moral philosophy often used to describe human behavior are *consequentialism* and *deontology*. Consequentialist theories posit that moral permissibility is evaluated solely with respect to the outcomes, and that one should choose the outcome with the highest value (Greene, 2007). On the other hand, deontological theories evaluate moral permissibility with respect to actions and whether they correspond to specific rules or rights.

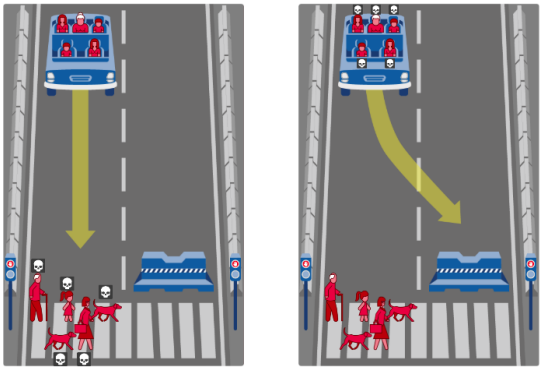
The trolley car dilemma (Foot, 2002; Thomson, 1984) highlights how these two families differ when making moral judgments. Here, participants must determine whether it is morally permissible to sacrifice an innocent bystander in order to prevent a trolley car from killing five railway workers. The "switch" scenario gives the participant the option to redirect the car to a track with one railway worker, whereas the "push" scenario requires the participant to push a large man directly in front of the car to stop it, killing the large man in the process. Given that the outcomes are the same for the "switch" and "push" scenarios (i.e., intervening results in one death, while not intervening results in five deaths), consequentialism prescribes intervention in both scenarios. Deontological theories allow for intervening in the "switch" scenario but not the "push" scenario because pushing a man to his death violates a moral principle, but switching the direction of a train does not.

Empirical studies have found that people are much more willing to "switch" than to "push" (Greene et al., 2001; Cushman et al., 2006), suggesting deontological principles factor heavily in human moral decision-making. Yet, a deontological theory's lack of systematicity makes it difficult to evaluate as a model of moral judgment (Greene, 2017). What are the rules that people invoke, and how do they interact with one another when in conflict? Furthermore, how do they interact with consequentialist concerns? Would people that refuse to push a man to his death to save five railway workers still make the same decision and with the same level of confidence when there are a million railway workers? Any theory of human moral cognition needs to be able to model how participants trade off different consequentialist and deontological factors.

Moral Machine Paradigm As society anticipates autonomous cars roaming its streets in the near future, the trolley car dilemma has left the moral philosophy classroom and entered into national policy conversations. A group of researchers aiming to gauge public opinion created "Moral Machine," an online game that presents users with moral dilemmas (see Figure 2) centered around autonomous cars (Awad et al., 2018). Comprising roughly forty million decisions from users in over two hundred countries, the Moral Machine experiment is the largest public dataset collection on human moral judgment.



(a) An autonomous car is headed towards a group of three pedestrians who are illegally crossing the street. The car can either stay and kill these pedestrians or swerve and kill three other pedestrians crossing legally.



(b) An autonomous car with five human passengers is headed towards a group of pedestrians who are illegally crossing the street. Staying on course will kill the pedestrians but save the passengers, while swerving will kill the passengers but save the pedestrians.

Figure 2: Two sample dilemmas in the Moral Machine dataset. In every scenario, the participant is asked to choose whether to *stay* or *swerve* (Awad et al., 2018).

In addition to the large number of decisions, the experiment operated over a rich problem space. Twenty unique agent types (e.g. man, girl, dog) along with contextual information (e.g. crossing signals) enabled researchers to measure the outcomes of nine manipulations: action versus inaction, passengers versus pedestrians, males versus females, fat versus fit, low status versus high status, lawful versus unlawful, elderly versus young, more lives saved versus less, and humans versus pets. The coverage and density of this problem space provides the opportunity to build a model that predicts how humans make moral judgments when a variety of different principles are at play.

Predicting Moral Decisions

As described earlier, the iterative refinement method we propose begins with both an initial, interpretable model and a more predictive black-box algorithm. In this section, we do exactly this by contrasting rational choice models derived from moral philosophy with multilayer feedforward neural networks.

Model Descriptions

We restricted our analysis to a subset of the dataset ($N = 12,478,340$) where an empty autonomous vehicle must decide between saving the pedestrians on the left or right side of the road (see Figure 2a for an example). The models we consider below are tasked to predict the probability of choosing to save the left side.

Interpretable Models Choice models (CM) are ubiquitous in both psychology and economics, and they form the basis of our interpretable model in this paper (Luce, 1959; McFadden et al., 1973). In particular, we assume that participants construct the values for both sides, i.e., v_{left} and v_{right}, and choose to save the left side when $v_{\text{left}} > v_{\text{right}}$, and vice versa. The value of each side is determined by aggregating the utilities of all its agents:

$$v_{\text{side}} = \sum_i u_i l_i \tag{1}$$

where u_i is the utility given to agent i and l_i is a binary indicator of agent i's presence on the given side.

McFadden et al. (1973) proved that if individual variation around this aggregate utility follows a Weibull distribution, the probability that v_{left} is optimal is consistent with the exponentiated Luce choice rule used in psychology, i.e.,

$$P(v_{\text{left}} > v_{\text{right}}) = P(c = \text{left} | v_{\text{left}}, v_{\text{right}}) = \frac{e^{v_{\text{left}}}}{e^{v_{\text{left}}} + e^{v_{\text{right}}}} \tag{2}$$

In practice, we can implement this formalization by using logistic regression to infer the utility vector $\mathbf{u}$. We built three models, each of which provided top-down different constraints on the utility vector. Our first model, "Equal Weight," required each agent to be equally weighted. At the other extreme, our "Utilitarian" model had no restriction. A third model, "Animals vs. People," was a hybrid: all humans were were weighted equally and all animals were weighted equally, but humans and animals could be weighted differently.

Research in moral psychology and philosophy has found that humans use moral principles in addition to standard utilitarian reasoning when choosing between options (Quinn, 1989; Spranca et al., 1991; Mikhail, 2002; Royzman & Baron, 2002; Baron & Ritov, 2004; Cushman et al., 2006). For example, one principle may be that allowing harm is more permissible than doing harm (Woollard & Howard-Snyder, 2016). In order to incorporate these principles, we moved beyond utilitarian-based choice models by expanding the definition of a side's value:

$$v_{\text{side}} = \sum_i u_i l_i + \sum_m \lambda_m f_m \tag{3}$$

where f_m is an indicator variable of whether principle m is present on the side and λ_m represents the importance of principle m. We built an "Expanded" model that introduces two principles potentially relevant in the Moral Machine dataset. The first is a preference for allowing harm over doing harm, thus penalizing sides that require the car to swerve in order to save them. Another potentially relevant principle is that it is more justified to punish unlawful pedestrians than lawful ones because they knowingly waived their rights when crossing illegally (Nino, 1983). This model was trained on the dataset to infer the values of $\mathbf{u}$ and λ.

Neural Networks We use relatively expressive multilayer feedforward neural networks (NN) to provide an estimate of the level of performance that statistical models can achieve in this domain. These networks were given as inputs the forty-two variables that uniquely defined a dilemma to each participant: twenty for the characters on the left side, twenty for the characters on the right side, one for the side of the car, and one for the crossing signal status. These are the same inputs for the "Expanded" choice model. However, the "Expanded" model had the added restriction that the side did not change an agent's utility (e.g., a girl on the left side has the same utility as a girl on the right side), while the neural network had no such restriction.

The networks were trained to minimize the crossentropy between the model's output and human binary decisions. The final layer of the neural networks is similar to the choice model in that it is constructing the value of each side by weighting different features. However, in these networks, the principles are learned from the nonlinear interactions of multiple layers and the indicators are probabilistic rather than deterministic.

To find the optimal hyperparameters, we conducted a grid search, varying the number of hidden layers, the number of hidden neurons, and the batch size. All networks used the same ReLU activation function and and no dropout. Given that most of these models both performed similarly and showed a clear improvement over simple choice models, we did not conduct a more extensive hyperparameter search. A neural network with three 32-unit hidden layers was used for all the analyses in this paper.

Model Comparisons

Standard Metrics Table 1 displays the results of the four rational choice models and the best performing neural network. All models were trained on eighty percent of the dataset, and the reported results reflect the performance on the held-out twenty percent. We report accuracy and area under the curve (AUC), two standard metrics for evaluating classification models. We also calculate the normalized Akaike information criterion (AIC), a metric for model comparison that integrates a model's predictive power and simplicity. All metrics resulted in the same expected ranking of models: Neural Network, Expanded, Utilitarian, Animals vs. People, Equal Weight.

Table 1: Comparison of Standard Metrics

Model Type	Accuracy	AUC	AIC
Equal Weight	0.571	0.616	1.301
Animals vs. People	0.630	0.702	1.234
Utilitarian	0.732	0.780	1.146
Expanded	0.763	0.826	1.046
Neural Network	0.774	0.845	0.983

Performance as a Function of Dataset Size Table 1 demonstrates that our cognitive models aren't as predictive as a powerful learning algorithm. This result, however, is only observable with larger datasets. Figure 3 plots each metric for each model over a large range of dataset sizes. Choice models performed very well at dataset sizes comparable to that of a large laboratory experiment. Conversely, neural networks improved with larger dataset sizes until reaching an asymptote where $N > 100,000$, at which point they outperform rational choice models. These results suggest that while psychological models are robust in the face of small datasets, they need to be evaluated on much larger ones.

Identifying Explanatory Principles

The neural network gives us an aspirational standard of how our simpler model should perform. Next, our task is to identify the emergent features it constructs and incorporate them into our simple choice model.

Calculating Residuals in Problem Aggregates By aggregating decisions for each dilemma, we can determine the empirical "difficulty" of each dilemma and whether our models predict this difficulty. For example, assume dilemmas A and B have been proposed to one hundred participants. If ninety participants exposed to dilemma A chose to save the left side and sixty participants exposed to dilemma B did, the empirical percentages for A and B would be 0.90 and 0.60, respectively. An accurate model of moral judgment should not only reflect the binary responses but also the confidence behind those responses.

We identified the specific problems where the neural network excelled compared to the "Expanded" rational choice model. Manually inspecting these problems and clustering them into groups revealed useful features beyond those employed in the choice model that the neural network is constructing. We formalized these features as principles and incorporated them into the choice model to improve prediction. Two examples are represented in Table 2.

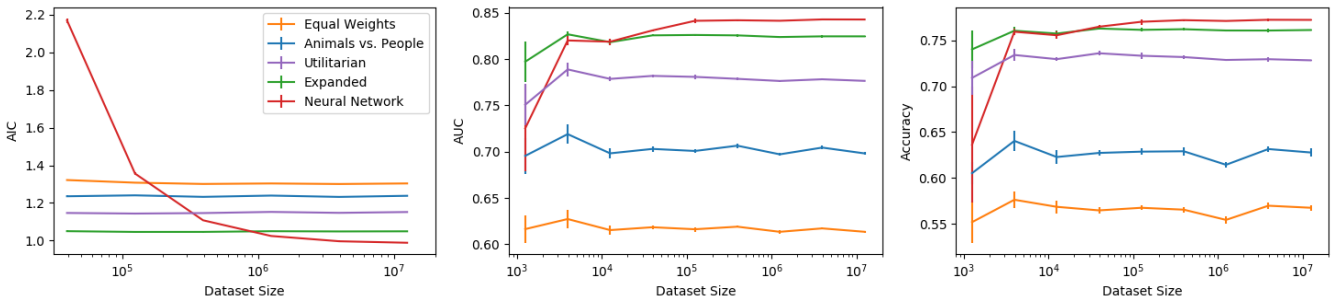
Table 2a describes a set of scenarios where one human is crossing illegally and one pet is crossing legally. Empirically, users tend to overwhelmingly prefer saving the human, while the choice model predicts the opposite. Our choice model's inferred utilities and importance values reveal a strong penalty (i.e., a large negative coefficient) for (1) humans crossing illegally and (2) requiring the car to swerve. However, the empirical data suggests that these principles are outweighed by the fact that this is a humans-versus-animals dilemma, and that humans should be preferred despite the crossing or intervention status. Thus, the next iteration of our model should incorporate a binary variable signifying whether this is an explicit humans-versus-animals dilemma.

We can conduct a similar analysis for the set of scenarios in Table 2b. Both models output significantly different decision probabilities, the neural network being the more accurate of the two. Most salient to us was an effect of age. Specifically, when the principal difference between the two sides is age, both boys and girls should be saved at a much higher rate, and information about their crossing and intervention status is less relevant. To capture this fact, we can incorporate another binary variable signifying whether the only difference between the agents on each side is age.

Incorporating New Features The two features we identified are a subset of six "problem types" the Moral Machine researchers used in their experiment: humans versus animals, old versus young, more versus less, fat versus fit, male versus female, and high status versus low status. These types were not revealed to the participants, but the residuals we inspected suggest that participants were constructing them from the raw features and then factoring them into their decisions.

Incorporating these six new features as principles resulted in 77.1% accuracy, nearly closing the gap entirely between our choice model and neural network performance reported in Table 1. Figure 4 illustrates the effects of incorporating the problem types into both the choice model and the neural network in details. Importantly, we observe that "Neural Network + Types" outperforms "Neural Network" at smaller dataset sizes, but performs identically at larger dataset sizes. This result suggests that the regular "Neural Network" is constructing the problem types we identified as emergent features given sufficient data to learn them from. More importantly, our augmented choice model now rivals the neural network's predictive power. And yet, by virtue of it being a rational choice model with only a few more parameters than our "Expanded" (and even the "Utilitarian") model, it remains conceptually simple. Thus, we have arrived at an interpretable statistical model that can both quantify the effects of utilitarian calculations and moral principles and predict human moral judgment over a large problem space.

Figure 4b still displays a gap between the AUC curves, suggesting there is more to be gained by repeating the process and potentially identifying new even more principles. For example, the last iteration found that when there was a humans-versus-animals problem, humans should be strongly favored. However, residuals suggest that participants don't honor this principle when all the humans are criminals. Rather, in these cases, participants may favor the animals or prefer the criminal by only a small margin. Thus, our next iteration will include a feature corresponding to whether all the humans are criminals. Our model also underperforms by overweighting



(a) Dataset Size vs. AIC (b) Dataset Size vs. AUC (c) Dataset Size vs. Accuracy

Figure 3: Test-set performance metrics of choice models and neural network[1] as a function of dataset size. Models were trained on five 80/20 training/test splits. Error bars indicate ± 1 SEM.

Table 2: Problem Aggregate Comparisons (Left Side Save Percentage)

Left Side Agents	**Right Side Agents**	**Car Side**	**Empirical**	**CM**	**NN**
Pregnant Woman Crossing Illegally	Cat Crossing Legally	Left	0.779	0.411	0.797
Stroller Crossing Illegally	Cat Crossing Legally	Left	0.826	0.425	0.801
Dog Crossing Legally	Male Doctor Crossing Illegally	Right	0.312	0.693	0.293
Cat Crossing Legally	Man Crossing Illegally	Right	0.308	0.692	0.266
Old Woman Crossing Illegally	Cat Crossing Legally	Left	0.670	0.306	0.622

(a) Problems indicating Human vs. Animals Principle

Left Side Agents	**Right Side Agents**	**Car Side**	**Empirical**	**CM**	**NN**
Old Man Crossing Legally	Boy Crossing Illegally	Right	0.350	0.647	0.341
Old Woman Crossing Legally	Girl Crossing Illegally	Right	0.337	0.642	0.321
Man	Boy	Left	0.113	0.417	0.097
Old Woman Crossing Legally	Girl Crossing Illegally	Left	0.268	0.570	0.269
Old Woman	Woman	Right	0.256	0.475	0.269

(b) Problems indicating Old vs. Young Principle

the effects of intervention. In problem types such as male versus female and fat versus fit, the intervention variable is weighted much differently than in young-versus-old dilemmas. The next iteration of the model should also include this interaction. Thus, this methodology allows us to continuously build on top of the new features we identify.

Conclusion

Large-scale behavioral datasets have the potential to revolutionize cognitive science (Griffiths, 2015), and while data science approaches have traditionally used them to predict behavior, they can additionally help cognitive scientists construct explanations of the given behavior.

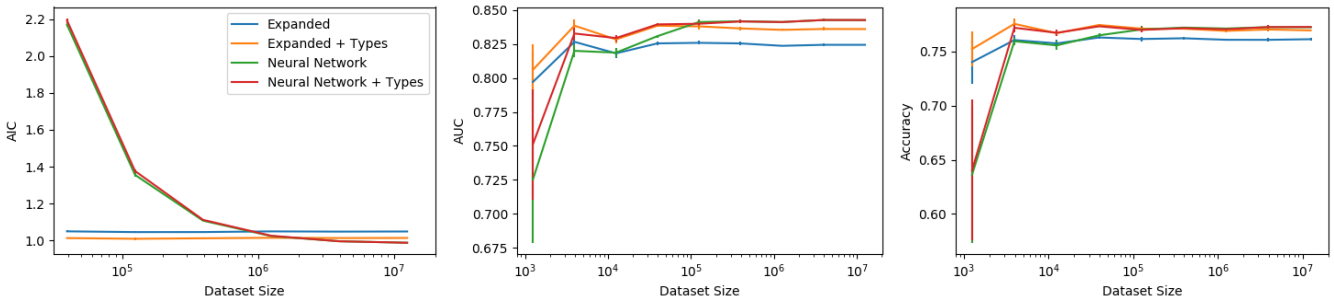
Black-box machine learning algorithms give us a sense of the predictive capabilities of our scientific theories, and we outline a methodology that uses them to help cognitive models reach these capabilities:

1. Amass a large-scale behavioral dataset that encompasses a large problem space
2. Formalize interpretable theories into parameterizable psychological models whose predictions can be evaluated
3. Compare these models to more accurate, but less interpretable black-box models (e.g., deep neural networks, random forests, etc.)
4. Identify types of problems where the black-box models outperform the simpler models
5. Formalize these problem types into features and incorporate them into both the simple and complex models
6. Return to Step 4 and repeat

We applied this procedure to moral decision-making, starting off with a rational choice model and iteratively adding principles until it had a comparable predictive power with black-box algorithms. This model allowed us to quantitatively predict the interactions between different utilitarian concerns and moral principles. Furthermore, our results regarding problem types suggest that moral judgment can be better predicted by incorporating alignable differences in similarity judgments (Tversky & Simonson, 1993), such as whether the dilemma is humans-versus-animals or old-versus-young.

The present case study, while successful, is only a limited application of the methodology we espouse, and further demonstrations are required to illustrate its utility. It will be

[1]While a batch size of 8,192 was used for Table 1, a batch size of 512 was used here because of the smaller dataset sizes.



(a) Dataset Size vs. AIC (b) Dataset Size vs. AUC (c) Dataset Size vs. Accuracy

Figure 4: Test-set performance metrics before and after incorporating new principles. Models were trained on five 80/20 training/test splits. Error bars indicate ± 1 SEM.

particularly interesting to apply our method to problems with even larger gaps between classic theories and data-driven predictive models. It is also likely that transferring insights from data-driven models will require moving beyond the sorts of featurization we consider here (i.e., problem clustering). In any case, we hope the microcosm presented here will inspire similarly synergistic approaches in other areas of psychology.

Acknowledgments. We thank Edmond Awad for providing guidance on navigating the Moral Machine dataset.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, *9*(1), 147–169.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... others (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, *1*, 203–232.

Box, G. E., & Hunter, W. G. (1962). A useful method for model-building. *Technometrics*, *4*(3), 301–318.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford handbook of moral psychology*, 47–71.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082–1089.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Foot, P. (2002). The problem of abortion and the doctrine of the double effect. *Virtues and Vices and Other Essays in Moral Philosophy*, 1932.

Greene, J. D. (2007). The secret joke of kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The neuroscience of morality: Emotion, brain disorders, and development* (Vol. 3, chap. 2). MIT Press.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Huang, Y., Cheng, Y., Chen, D., Lee, H., Ngiam, J., Le, Q. V., & Chen, Z. (2018). Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint arXiv:1811.06965*.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? *arXiv preprint arXiv:1604.02416*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Linderman, S. W., & Gershman, S. J. (2017). Using computational theory to constrain statistical models of neural data. *Current opinion in neurobiology*, *46*, 14–24.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*.

Lzaro-Gredilla, M., Lin, D., Guntupalli, J. S., & George, D. (2019). Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. *Science Robotics*, *4*(26).

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Nino, C. S. (1983). A consensual theory of punishment. *Philosophy & Public Affairs*, 289–306.

Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of doing and allowing. *The Philosophical Review*, *98*(3), 287–312.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of experimental social psychology*, *27*(1), 76–105.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review*, *88*(2), 135.

Thomson, J. J. (1984). The trolley problem. *Yale Law Journal*, *94*, 1395.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management science*, *39*(10), 1179–1189.

Woollard, F., & Howard-Snyder, F. (2016). Doing vs. allowing harm. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.