COG403 Moral Machine

Ezra Robens-Paradise

Elif Erzincan

University of Toronto, Department of Cognitive Science

Abstract

This paper explores the integration of moral decision-making in artificial intelligence (AI), emphasizing the need for AI systems to align with human ethical standards. As AI becomes increasingly autonomous, the concept of artificial moral agents (AMAs) capable of ethical reasoning, in a manner that is similar to humans, becomes crucial. The paper introduces a computational model using data from the Moral Machine experiment to replicate human moral judgments that pertain to life-and-death scenarios. By incorporating factors like age, fitness, and social status, the model aims to assess the value of human life as humans do. The findings reveal that characteristics such as gender, age, and social status significantly influence moral decisions, highlighting the complexity of human ethics. The computational model also has implications on human moral cognition, as it offers a structural account of moral decision-making.

Keywords: AI alignment; moral cognition; moral machine; ethical decision-making

Introduction

With the surge of AI tools being implemented in all aspects of society, it is important that our tools can assess the value of human life. AI alignment is about making sure that artificial intelligence systems behave in ways that are consistent with human values and goals. Wallach and Wallor (2020) define artificial moral agents (AMAs) as computer systems or robots that can make moral decisions or act in morally sensitive ways and argue that AMAs are necessary and inevitable as autonomous systems become more pervasive and sophisticated and that they pose both technical and ethical difficulties for engineers and philosophers. In this context, the establishment of a system to accurately determine the value of human life, more specifically how the value of two people's lives compare, becomes not just an abstract philosophical question, but a practical imperative. This is critical for any AI system that has a degree of agency and the potential to affect people's lives, which includes but is not limited to cars, triage algorithms, and automated weapons.

At the same time, human beings are not equipped to tackle modern large-scale moral challenges like climate change, and moral AI may help humanity overcome this evolutionary limitation. Similarly, although biases in moral judgements are surpassed after systematic moral reasoning, most moral judgements are based on instant emotions and intuitions about a situation (Savulescu & Maslen, 2015). Thus, moral AI agents may play a significant role in helping agents overcome their biases in making moral judgements. However, strong moral AI would call upon creating an agent that replicates and exhibits the best and unbiased moral qualities of humans. Savulescu and Maslen (2015) suggest that another option is a kind of weak moral AI that assists humans in their moral decision-making processes.

While Johnson's human agency approach maintains that only human beings can be moral agents, and that AI systems are merely tools that extend human capabilities and responsibilities, Floridi suggests that AI systems can also be considered moral agents, but with a different set of criteria. Floridi's artificial agency approach argues that instead of limiting moral agency to entities that have intentions and mental abilities like humans, AI systems that are interactive, autonomous, and adaptive should also be considered moral agents (Fritz et al., 2020). A categorization that bridges these differing viewpoints comes from Moore (2006) who defines "full ethical agents" as entities that have consciousness, intentionality and full autonomy over actions. He suggests that so far, machines have reached the level of "explicit ethical agency" which allows them to reason about ethics (Winfield et al., 2019). Moore's concept of explicit ethical agency offers a pragmatic solution for the time being: This notion accepts that, although current AI systems lack consciousness and intentionality, which are key components of full ethical agency, their ability to reason about ethical matters is a significant step forward. In light of this, our focus will be on leveraging and enhancing this explicit ethical agency in AI. Shaw et al. (2018) list four principles that must be present in a moral agent: Robustness, consistency, universality, and simplicity. A moral agent must be adaptable and capable of changing its moral principles based on different contexts. Moral agents should be able to adopt the moral laws of any society and learn new principles when transplanted to a different cultural environment. Regardless of the specific moral principles learned, they should be internally consistent, and the agent's moral rules should not contradict each other. The moral principles learned by the agent should be universally applicable to all members of its society. They should not be overly restrictive or arbitrary. While having a comprehensive set of moral principles is important, a moral agent should strive to operate with the smallest number of "firm" moral principles possible. This avoids sacrificing diversity for homogeneity and promotes efficiency (Shaw et al., 2018). In attaining a model of morality, these will be the guiding meta-qualities we want our moral system to have.

On the other hand, mechanizing how humans form moral judgements presupposes their formalization, which is the central point of moral philosophy debates for many years. To attain ethical autonomous systems that take actions based on ethical value and implicit ethical agents that avoid unethical outcomes, the cognitive architecture that underlies ethical choices must be understood and realized in machines (Winfield et al., 2019). By doing so, we can better understand how humans process moral and ethical scenarios, which is a key component of human cognition. It also allows a more structured way to analyze and assess ethical theories of moral philosophy and compare different moral frameworks like utilitarianism, deontology, or virtue ethics. However, one challenge in designing an ethical autonomous system is that it requires the ability to acknowledge the ethical salience of a given action (Winfield et al., 2019). Moreover, creating moral AI implies specifying human moral norms, characterizing human moral cognition, and choosing the technical approach equipped to capture the complex nature of human morality (Haas, 2020). Particularly, knowledge-based models of morality would require a fleshed-out encoding for normative principles and moral context, which demands significant effort and consideration of cultural and contextual differences in judgements of morality.

The challenges of AI alignment and the imperative to formalize human moral judgments, as discussed, are foundational to the ethos of our project. The motivation of this project is to establish a potential computational model that aims to replicate humans' ability to assign human value to others. Such a model is pivotal for AI systems tasked with making decisions that could significantly impact human lives, ensuring that these decisions are aligned with a well-understood and formally structured interpretation of human ethics and morality. To do so, we will use the data from The Moral Machine experiment, which is a project that asks users to make choices in scenarios involving self-driving cars that are going to crash (Jaques, 2019; Awad et al., 2018). We hypothesize that by utilizing a computational model that incorporates factors such as fitness level, age, species, and social status, we can develop a system that accurately replicates human moral judgments about individual worth using a back propagation neural network. This system will be able to replicate how humans attempt to assess the value of human life in ethical dilemmas, thereby contributing to the field of AI alignment and ensuring that artificial intelligence systems behave in ways that align with human values. At the same time, we think that the results of the analysis may yield interesting results about the framework of human moral cognition.

Ethical Disclaimer

This paper explores the development and use of a moral decision-making AI that assigns value to specific characters. This model should not be used to inform real-world moral decisions nor are the conclusions of the paper representative of objective values of human life. The purpose of this paper is academic exploration and an attempt to model human moral reasoning.

Methods

To provide a comprehensive understanding of our methodology and to ensure the reproducibility of our results, we include below a link to the code used in our analysis: `https://github.com/ezrarp/moral_machine.git` for help navigating the repository see the README.

Data Cleaning

In this data set participants are presented with a series of hypothetical situations where a self-driving car must choose between causing harm to different groups of people. Each scenario involves a collision, and the participant must decide who the car should prioritize to minimize harm. This dataset consists of 41 columns. The final 20 columns indicate the count of characters of various types in each outcome. For instance, the "Man" and "Woman" columns denote the number of Man and Woman characters in each outcome. In this dataset, the Man column signifies the frequency of the character Man's presence in that outcome, rather than the total occurrences of male characters. 20 different characters include pets, people of higher social status, people of different health statuses and differing ages. Each set of two rows represents a situation with each of the rows describing who would be killed in that outcome. Each situation contains the response that was chosen by the human participant.

The data set consists of approximately 3.9 million scenario decisions made by participants. Due to the vast size of the data set and our limited computational access, we opted to take a random sample of the data. We randomly selected 350,000 scenarios to be included in the training of the model and an additional random sample of 5000 scenarios for testing. The data cleaning process involved splitting the sampled sub-dataset into three parts: Two datasets, right-hand side and left-hand side, contain character information from the original dataset, where the existence of a character is quinary. Each corresponding row in the right-hand side and left-hand side datasets make up pairs of scenarios that participants were asked to choose from. The targets dataset stores the ground-truth, which is a binary variable with value 1 when the participant chose to save the characters in the scenario stored in the right hand side, and 0 otherwise.

Logistic Regression

We used the sklearn Python package to build a logistic regression model that uses the difference between character features within two scenarios to predict which would be chosen to be saved. This approach helps in understanding which character features and to what extent they influence the choice between two scenarios. The difference essentially captures the relative importance or weight of each character in deciding between the two scenarios. The explanatory variables included the difference between values of the following characters in each pair of scenarios: 'Man', 'Woman', 'Pregnant', 'Stroller', 'OldMan', 'OldWoman', 'Boy', 'Girl', 'Homeless', 'LargeWoman', 'LargeMan', 'Criminal', 'MaleExecutive', 'FemaleExecutive', 'FemaleAthlete', 'MaleAthlete', 'FemaleDoctor', 'MaleDoctor', 'Dog', 'Cat'. This means we subtracted the value of the character feature in one scenario from its value in the other scenario. The response variable was the 'Saved' value of the scenario that we subtracted the character features from. To ensure that all features contribute equally to the model, we scaled them using the Standard Scaler, which standardizes features by removing the mean and scaling to unit variance. The dataset was split into training and testing sets, with 70% of the data being used for training, and 30% for testing the performance of the model. An L2 regularization of 0.1 was applied, to prevent overfitting and penalize large coefficients. A coefficient analysis was conducted.

Neural Network

The Neural Network was implemented using the Pytorch library in Python, it was a 5-layer supervised learning model that used back-propagation for training. The model had an input size of 20 and subsequent layers consisting of 16, 9, and 3 nodes, with 1 output node. The final node had an output classifier function that classified the output as 1 if the input was greater than 0.5 and 0 if the input was less than 0.5. The network was fully connected and initialized with random weights. We used the sigmoid activation function ($f(x) = \frac{1}{1+e^{(-x)}}$) for each node. The input consisted of the vector difference between the left input and the right input ($Input = x1 - x2$) where each vector was the number of characters that would be killed on each side. The architecture of the model can be seen in Figure 1

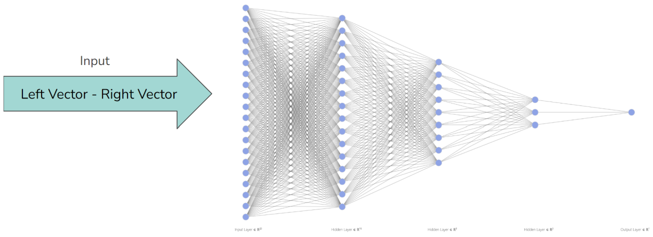


Figure 1: Model Architecture

The model was trained on 350,000 inputs of vector differences for 10 epochs. We used Google Collab's T4 GPU. This was a supervised learning model, where the target values consisted of the decision that was made by the participant who answered the scenario. The model used Mean-Squared error as the loss function ($MSE = \sum \frac{(y_i - m_i)^2}{n}$), and using Stochastic-Gradient Descent as an optimizer, the loss was back-propagated across the model with a learning rate of 0.01. The model was then tested for accuracy against a unique sample data set that was not part of the training data with a length of 5000.

The output of the model before the classification function is the comparative moral value (M_c) of each character vector. A character vector could be compared to an empty vector to determine its raw moral value (M_r).This was used to determine each character's individual moral ranking. In this process, the first character vector (v_1) is subtracted from the second character vector (v_2). The vector difference then passes through the network and the output represents v_2's M_c relative to v_1, then the process is symmetrically repeated where the input is $v_1 - v_2$ to provide us with v_1's M_c relative to v_2. These values can then be compared.

To evaluate the significance of features in the network's predictions, we performed a permutation feature analysis on the trained Pytorch model. We measured the model's prediction accuracy on the unaltered test data to be used as a baseline. Then, 1000 permutations were performed over the features in the right input. In each iteration, one feature at a time was randomly shuffled (permuted) in the test data while keeping the other features unchanged. This was done separately for both the right input and the left input first, and then for the difference of features between the right and the left input. After permuting a feature, the model's accuracy was recalculated. The drop in accuracy due to the permutation indicates the importance of the feature, as when the significant features are permuted, they will lead to a substantial decrease in model performance. Finally, the decrease in accuracy for each permuted feature was summed over all iterations. This accumulated sum represents the overall importance of each feature. The summed importance scores were then averaged over the total number of permutations to normalize them.

Results

Below is a summary of the results we obtained from the logistic regression model and the neural network. This includes performance evaluations, coefficient analysis, feature importance analysis, and the results of moral comparisons.

Logistic Regression

We found the accuracy of the logistic regression model to be 69.492%. The equation of the model is given in Figure 12 of Appendix.
Here are the results of the coefficient analysis of the model:

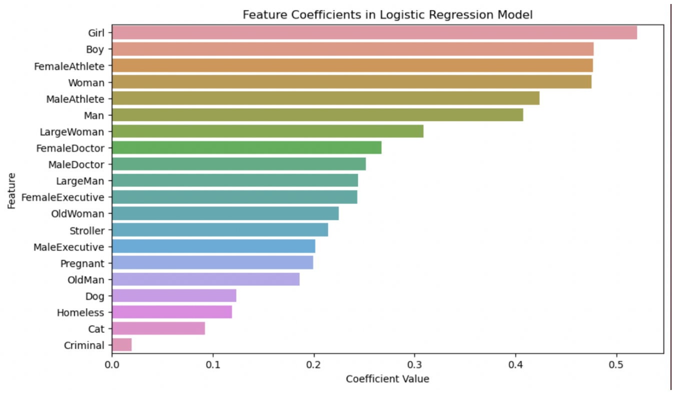


Figure 2: Coefficient Analysis

The above plot illustrates the relationship between feature coefficients within the logistic regression model and the log odds of the outcome occurring. A higher coefficient for a given feature suggests a stronger association with an increased likelihood of the decision to save the characters in the scenario, as per the model's prediction. Then, based on the results of the coefficient analysis we can infer that the top features with most influence on the model's predictions are the characters girl, boy, female athlete, and woman. This suggests that, all else being equal, scenarios with these characters are more likely to result in the characters being saved. The opposite is true for characters with the smallest coefficients, namely, criminal, cat, homeless, and dog. It's important to note that the values represent the log odds. Therefore, the influence of the features on the probability is not linear; small changes in log odds can result in larger changes in probability, especially when the log odds are near zero.

Neural Network

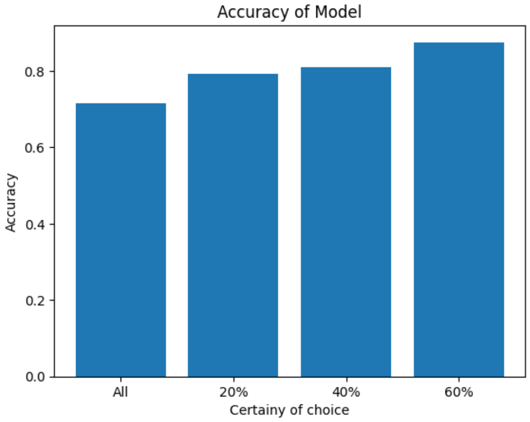


Figure 3: Model Accuracy Graph

Accuracy and Loss The model was able to make the same moral judgments as human respondents 71.6% of the time. This raw accuracy could be improved if we select only decisions where the model produced a high degree of difference in M_c between the two vectors. Limiting the decisions to only include outputs with a difference of $0.6M_c$ increased the model's accuracy to 87.5% as can be seen in Figure 3.

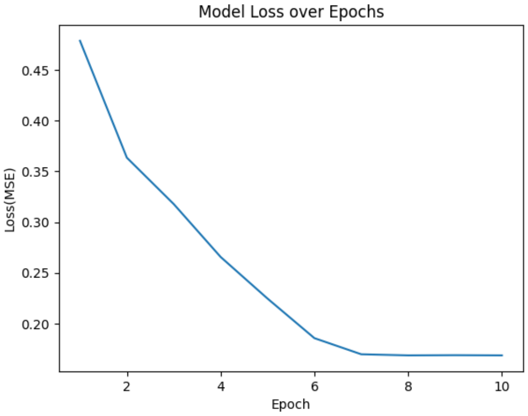


Figure 4: Model loss over epoch

In this scenario, we only tested the accuracy against results that had M_c less than 0.3 or greater than 0.6. This can also be described as the model's confidence value; the more extreme the M_c, the more confident the model is in its decision. The correlation between high M_c values and improved accuracy suggests that the model's confidence is a reliable indicator of the correctness of its moral judgments. Higher confidence levels are associated with more accurate decisions, reflecting a strong alignment with human moral reasoning.

When training the model the loss, calculated by the Mean-Squared Error appeared to plateau at approximately 0.168 after the seventh training epoch as seen in Figure 4.

Feature Importance The following plots are the feature importance analysis for the right input and the combined feature analysis, respectively.

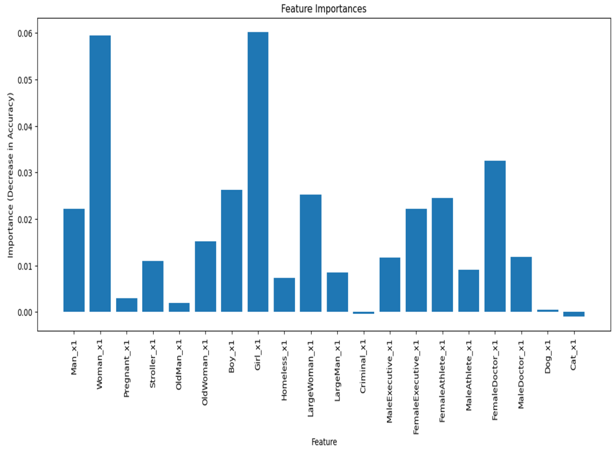


Figure 5: Feature Importance Analysis for the Right Input

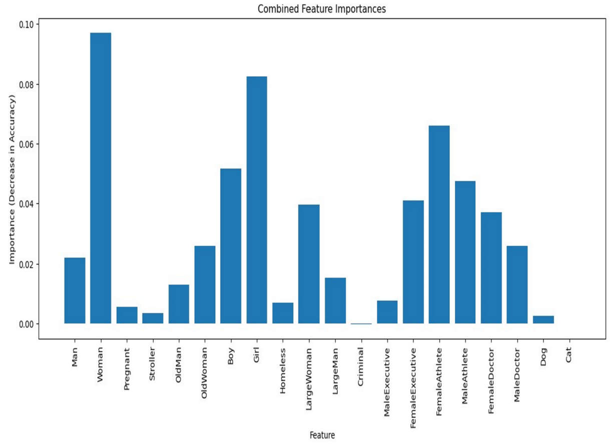


Figure 6: Combined Feature Importance Analysis

The two plots have similar distributions, which suggests that there is symmetry and robustness in decision-making, as the network treats the features from both sides similarly. In both plots, woman and girl are the most important features, which suggests that the network relied more heavily on these characters for making accurate predictions. Similarly, criminal, dog, and cat are the least important features in both plots, implying that they are less important for the network's predictions. These results align with the findings from the coefficient analysis for the logistic regression model, given in Figure 2.

Moral Ranks Each character was compared to an empty vector to determine its M_r. These values were then ordered to determine the model's moral rankings for each individual. The 'Stroller' character had the highest M_r ($M_r = 0.7523$). The 'Criminal' character had the lowest M_r ($M_r = 0.5107$). The base M_r is 0.5, implying moral indifference. The low M_r implies that criminals are only barely worth saving. All the character's rankings can be seen in Figure 7.

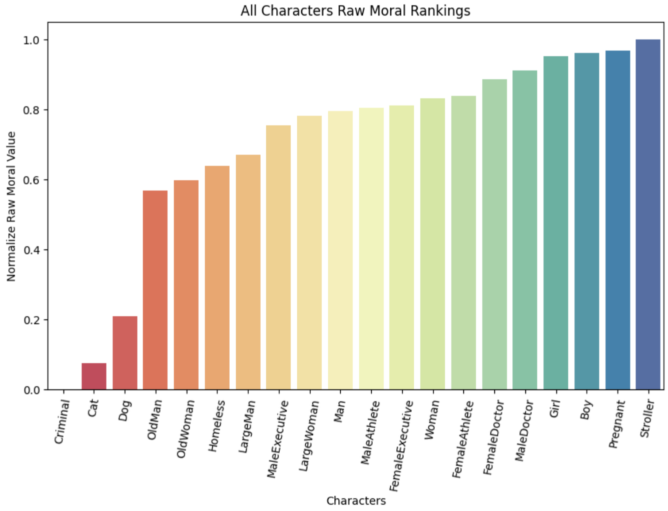


Figure 7: Raw Moral value of each character

The trained model allows us to create moral comparisons between character scenarios, v_1 and v_2, and output an M_c value. A $M_c > 0.5$ when the input is passed through $v_1 - v_2$ implies that the model prefers saving the characters from v_1. Using this method we were able to compare characters based on gender, age, fitness and species. Male and female characters had a compared M_c of 0.4908 for males, and 0.5157 for females, meaning that the model preferred saving female characters over males as seen in Figure 8. There is a significant preference to not save Large people compared to both fit and average people. The M_c of 'large' people was substantially less as seen in Figure 9. In the cross-species comparison, the model valued approximately 5 dogs are worth 1 man, with the model opting to save 5 dogs with a low degree of confidence ($\leq 10\%$) as seen in Figure 10. There was also a significant difference in M_c when comparing character age. Children have a significantly higher M_c compared to old characters as seen in in Figure 11.

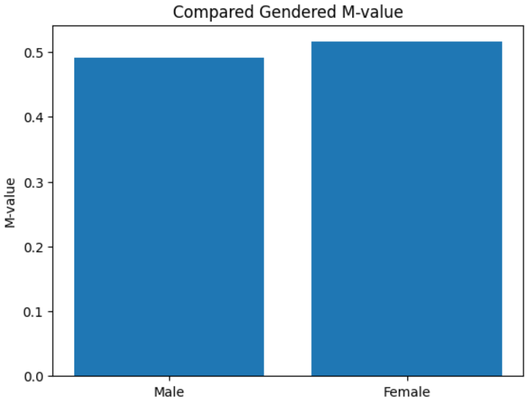


Figure 8: Compared Gender M_c

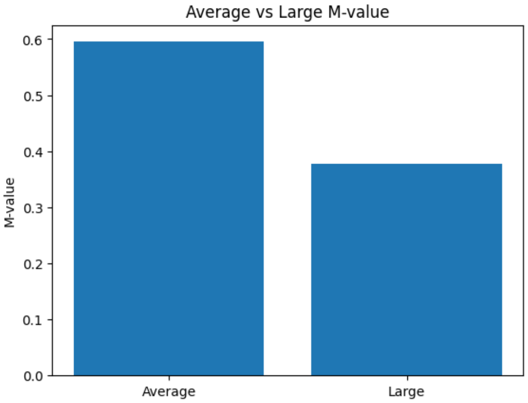


Figure 9: Large Characters vs Average Characters relative M_c

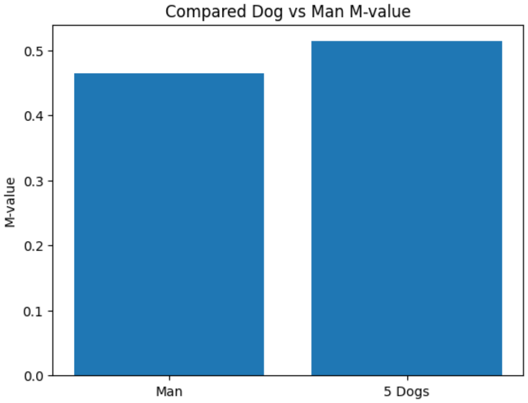


Figure 10: Number of dogs with equivalent M_c to a man

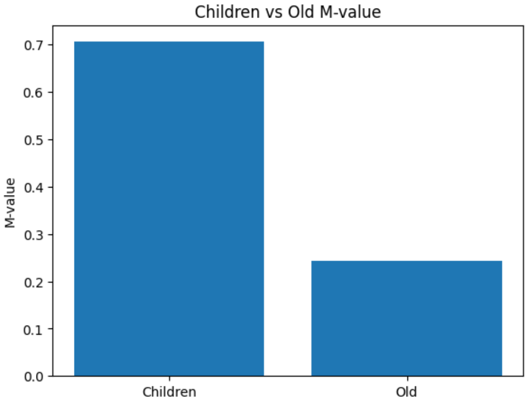


Figure 11: Young vs Old characters relative M_c

The comparative plots are the result of our eye test, that served as a sanity check for the model. This is to ensure that the moral outcomes of the model are qualitatively reasonable.

Conclusion

We hypothesized that we could model moral decision-making in humans using supervised learning. We formalized this cognitive process, producing a model replicating human moral cognition and assigning moral value to individuals. We saw that gender, age, fitness level, species and social status play key roles in our moral decision-making, with each feature holding a different significance.

Discussion

Omitted Data The data set contained columns that labelled if the characters were crossing the street illegally. We chose to omit this feature of the data to simplify the model and increase the interpretability. We also omitted the feature of the data that labelled the scenario as an action or inaction. This means that humans consider the difference between committing an action that resulted in death or doing nothing that resulted in death. It was found that humans tend to prefer inaction to action when making these moral judgements.

We chose to omit this feature as we wanted our model to behave based on a utilitarian moral model, where it chooses to maximize preserving M_c rather than considering a subjective personal cost to the model. We are confident that the omission of this feature of the data did not disrupt the symmetry of the model's decision-making process. The model consistently values inputs symmetrically in all of our comparison testing, where a vector comparison would have a high M_c, and the same comparison with swapped input would have an equivalently low M_c. This claim is additionally supported by the consistency of symmetry present in the feature analysis.

Neural Network Benefits and Cognitive Significance Our decision to build a neural network is a consequence of the relatively limited accuracy of our baseline logistic regression model, which may potentially be due to the data not being linear. Moreover, the hierarchical feature recognition and representation properties of the neural network are similar to how humans assess moral information at different levels, thus providing a more sophisticated and nuanced approach to modelling moral decision-making that resembles the structure of human moral cognition. This approach aligns with the emergent understanding in cognitive science that human decision-making involves an intricate web of value assessments, emotional influences, and contextual considerations—all of which can be more accurately captured with the flexible architecture of neural networks than with models assuming linearity, such as logistic regression.

Limitations

One limitation of our project stems from the inherent complexity of human morality, which varies widely across cultures, personal experiences, emotional states, and societal norms (Awad et al., 2018). People tend to shape their moral beliefs based on the influence of these factors, but our models do not account for these multifaceted influences. Consequently, this limitation may affect the generalizability and applicability of our findings across different demographic and cultural contexts. Similarly, ethical dilemmas in the real world often involve nuanced and complex situations that go beyond the scope of the dataset. As discussed by Savulescu & Maslen (2015), immediate emotions and intuitions significantly influence moral judgments, but these subtleties are not captured by our dataset, which may limit the depth of our analysis in mirroring real-world moral complexity. At the same time, even if we remedied these limitations, the biggest challenge in the practical use of such tools remains to be the ethical responsibility about relying on computational models for ethical decision-making.

Further Research

To attain improved accuracy and overcome the limitations we listed, we think further research should experiment with more advanced neural network architectures and focus on accounting for the longitudinal and cross-cultural effects of morality. Moreover, research on the broader ethical and societal implications of delegating moral decision-making to AI is much needed, to resolve the fundamental concern of entrusting ethical decision-making to computational models.

References

Awad, E., Dsouza, S., Kim, R. et al. (2018). *The Moral Machine experiment.* Nature 563, 59–64. https://doi.org/10.1038/s41586-018-0637-6

Fritz, A., Brandt, W., Gimpel, H., & Bayer, S. (2020). *Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI).* De Ethica, 6(1), 3-22.

Haas, J. (2020). *Moral Gridworlds: A Theoretical Proposal for Modeling Artificial Moral Cognition.* Minds and Machines (Dordrecht), 30(2), 219–246. https://doi.org/10.1007/s11023-020-09524-9

Jaques, A. E. (2019). *Why the moral machine is a monster.* University of Miami School of Law, 10, 1-10.

Moor, J. H. (2006). *The nature, importance, and difficulty of machine ethics,* IEEE Intell. Syst., vol. 21, no. 4, pp. 18–21, Jul./Aug. 2006.

Savulescu, J., & Maslen, H. (2015). *Moral Enhancement and Artificial Intelligence: Moral AI?.* In: Romportl, J., Zackova, E., & Kelemen, J. (eds) Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics, vol 9. Springer, Cham. https://doi.org/10.1007/978-3-319-09668-1_6

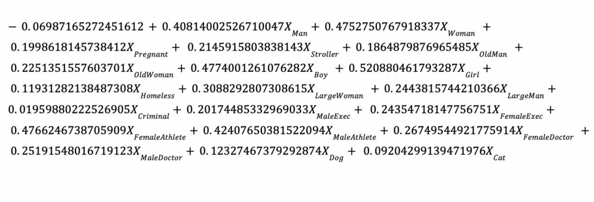
Shaw, N. P., Stöckel, A., Orr, R. W., Lidbetter, T. F., & Cohen, R. (2018, December). *Towards provably moral AI agents in bottom-up learning frameworks.* In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 271-277).

Wallach, W., & Vallor, S. (2020). *Moral machines.* Ethics of Artificial Intelligence. Oxford University Press, 383-412.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). *Machine ethics: the design and governance of ethical AI and autonomous systems.* Proceedings of the IEEE, 107(3), 509–517.

Appendix

$P(\text{Saved} = 1) = \frac{1}{1+e^{-z}}$, where z is given below.



$$- 0.06987165272451612 + 0.40814002526710047X_{Man} + 0.4752750767918337X_{Woman} + 0.1998618145738412X_{Pregnant} + 0.2145915803838143X_{Stroller} + 0.1864879876965485X_{OldMan} + 0.225135155760370 1X_{OldWoman} + 0.4774001261076282X_{Boy} + 0.520880461793287X_{Girl} + 0.11931282138487308X_{Homeless} + 0.3088292807308615X_{LargeWoman} + 0.2443815744210366X_{LargeMan} + 0.01959880222526905X_{Criminal} + 0.20174485332969033X_{MaleExec} + 0.2435471814775 6751X_{FemaleExec} + 0.4766246738705909X_{FemaleAthlete} + 0.42407650381522094X_{MaleAthlete} + 0.26749544921775914X_{FemaleDoctor} + 0.25191548016719123X_{MaleDoctor} + 0.12327467379292874X_{Dog} + 0.09204299139471976X_{Cat}$$

Figure 12: Logistic Regression Equation

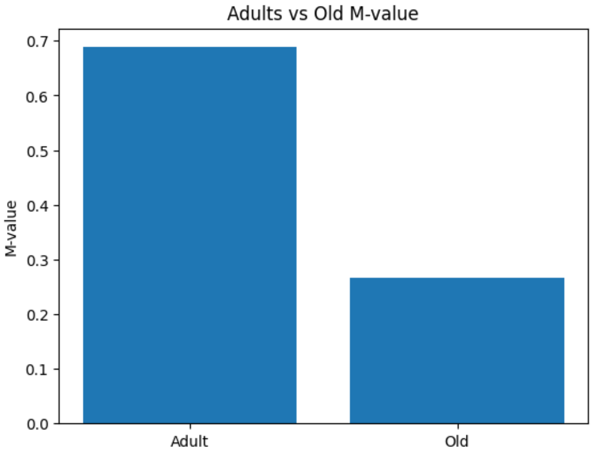


Figure 13: Adult character's comparative moral value to old characters

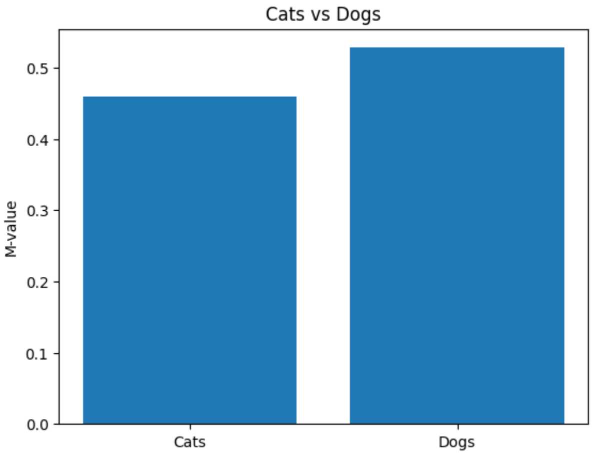


Figure 14: Cats and dogs comparative moral Value

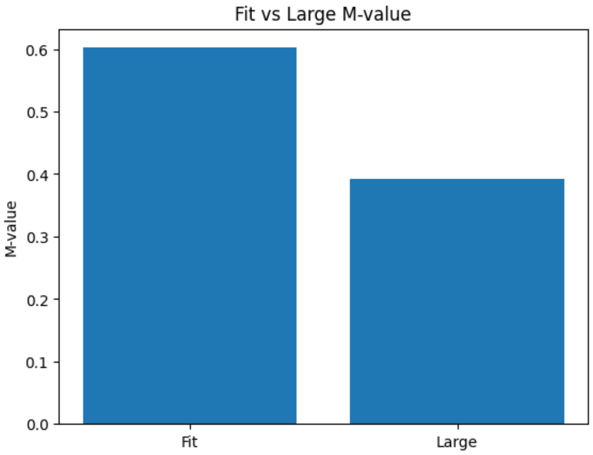


Figure 15: Fit characters and large character's comparative moral value

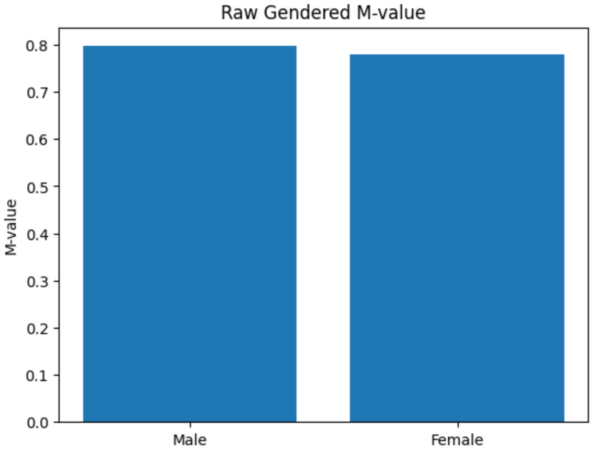


Figure 16: Gendered characters raw moral value