Milestone Report

Dept of Neuroscience and Psychology

Paul McKee, Sanskriti Purohit,

Natalie Smith, Kristi Van Meter

## **Introduction**

Numerous social interactions are subject to the sway of moral and racial attitudes, whether overt or covert. Human history is stained with the blood shed by harmful explicit racial attitudes such as the systematic genocide of approximately six million Jews (Holocaust), the mass slaughter of Tutsis by ethnic Hutus (Rwandan Genocide), and the ongoing detainment and forced sterilization of Uyghurs in China. While the damage caused by these explicit attitudes is clearly seen, the costs of implicit racial attitudes, although perhaps smaller in scale, are no less real and are paid in full by those affected. Studies have shown that teachers may hold lower academic expectations for students of color, which can lead to a self-fulfilling prophecy where these students perform worse than their potential (Jussim & Harber, 2005). Research has also demonstrated that healthcare providers may be more likely to under prescribe pain medication for Black patients compared to White patients, due to unfounded beliefs about pain tolerance or the likelihood of substance abuse (Hoffman et al., 2016).

While we may have moral judgments about the above topics, they are not inherently about morality in and of themselves. For example, we may judge someone as immoral for being sexist, but the act of sexism does not necessitate the perpetrator having moral judgments toward the victims in order to treat them differently (i.e., “Women are less morally worthy and should be treated as such”). In a similar vein, many people may agree it is moral to not hold racist attitudes; however, the decision to treat different races similarly may not necessarily involve moral cognition during the act. Racial and gender implicit attitudes can be considered moral implicit attitudes when their content is explicitly linked to moral evaluations, such as "Black people are morally worse than White people." That is, implicit moral attitudes are the automatic and unintentional evaluations in response to the perceived morality of an individual or situation. However, if the content of these attitudes reflects non-moral evaluations or beliefs, such as associating certain stereotypes with specific racial or gender groups without moral connotations, then they would not be considered moral implicit attitudes. The distinction between moral and non-moral implicit attitudes relies on the presence or absence of a moral evaluation in the content of the attitude itself.

There is considerable evidence that the intertwining of identity – such as race – and moral attitudes occurs (McKee et al, in review). Many studies have found a tendency to judge immoral acts done by friends or other ingroup members less harshly than strangers or outgroup members (Valdesolo & DeSteno, 2007; De Bock et al, 2013; Bandura, 2016; Forbes & Stellar, 2022). Research also supports the existence of a black sheep effect. This effect suggests that norm and value conforming in-group members tend to be judged more positively than comparable outgroup members. On the other hand, norm and value deviating in-group members tend to be judged more negatively than comparable outgroup members (Marques, Yzerbyt, & Leyens, 1988; Abrams et al, 2013; Bettache et al, 2019). While ingroups (vs. outgroups) don’t have to be racially defined (e.g., family, sports team), they often are.

In addition to the varying identity of the agent, the victim of an immoral act, or recipient of a moral act, can be either an ingroup member or an outgroup member. This varying identity might affect moral judgments of acts that harm or benefit that individual. One study speaking to the identity of the victim, found that defendants charged with killing white victims were 4.3 times as likely to receive a death sentence than defendants charged with killing black victims, even after accounting for 39 nonracial variables (Baldus et al. 1983; cited in McCleskey v. Kemp 1987). Though sentencing is distinct from moral judgment, this suggests that people might be less harsh in their moral judgments of wrongs when they see the victim as a member of an outgroup.

The objective of our study is to expand on our understanding of the manifestation of these attitudes and their impact on human behavior. We will investigate whether individuals exhibit an automatic behavioral manifestation in response to situations perceived as moral good, bad, or neutral that can shed insight into their implicitly held moral attitudes. We are doing so by studying the speed, and any reaction time asymmetries, when people are forced to approach or avoid various moral scenarios using an adapted approach-avoidance task. We will then examine how these approach-avoidance tendencies relate to implicit and explicit moral attitudes and also if these automatic tendencies change based on the race of the people involved in the morally relevant scenarios.

To gain a deeper understanding of our implicit moral attitudes, we will investigate the following research questions:

RQ1: Are there differences in the speeds in which people automatically approach or avoid stimuli of varying moral content?

RQ2: Does the race of the individuals involved in stimuli of varying moral content affect the speeds in which people automatically approach or avoid?

RQ3: Do these racial-moral automatic approach avoidance tendencies relate to implicit and explicit moral attitudes?

Despite its potential importance in our daily lives, no research has been done to understand the role of approach-avoidance motivation in our moral choices. Our approach-avoidance motivation is the automatic tendency to move towards things we like and away from things we don't. This project aims to shed light on this relationship. The potential benefits of understanding this are vast, from everyday decision-making to complex societal issues. One of the key outcomes could be a new perspective on understanding the motivations of individuals with certain criminal tendencies, like psychopaths. To give an example, if we find that psychopaths don't show a typical approach-avoidance response to certain moral scenarios, it might offer a clue into their unique mindset.

## **Methods**

The data used in this project was collected from an online experiment. The data consists of a study where the stimuli were three-word phrases and a study where they were images. An a priori power analysis (G\*Power) revealed a required sample size of 690 participants, in order to detect an effect size f = 0.15 (with α = 0.05 and 1 − β = 0.95) for each study. To maximize our chances and prevent possible data loss, we decided to collect slightly more participants. The participants (Phrases: N = 757; 50.5% female; mean age = 45.2, SD = 16.7; Images: N = 728; 49.7% female; mean age = 47.3, SD = 17.4) were recruited from Qualtrics and completed the online study for monetary compensation. The sample was diverse in regards to race and ethnicity.

**Materials and Procedure**

Participants started by completing demographic questions on a Qualtrics survey. If they were deemed eligible according to recruitment quotas, they were then randomly assigned into an experimental task that was completed on PsyToolKit.org. The stimuli for either of the studies were three-word phrases or images categorized into morally good, morally bad, or morally neutral. Each phrase or image was also presented with either a yellow or blue background. After determining eligibility, participants were randomly allocated to one of the primary between-subjects treatment arms: Implicit (Indirect) Stimulus Evaluation: Participants responded based on an irrelevant characteristic to the moral content - the color of the background (yellow or blue), or Explicit (Direct) Stimulus Evaluation: Participants responded based on the actual moral content of the stimuli. Subsequent to this primary allocation, participants were then randomly assigned to view two types of moral content from three possible combinations: morally good and morally neutral, morally good and morally bad, morally neutral and morally bad.

A further randomization was implemented to counter-balance the instructions provided. Specifically, In the Implicit Evaluation arm half were instructed to approach yellow stimuli and avoid blue while the other half were instructed vice-versa. In the Explicit Evaluation arm, for each moral content pairing, half of the participants were given instructions to approach one type of content and avoid the other. This was then reversed for the remaining participants.

During the task, participants used the “U” key to "approach" a stimulus, causing it to appear closer by enlarging it. Conversely, pressing “M” caused participants to “avoid” the stimulus, making it appear farther by reducing its size. Before the main trials, participants undertook twenty practice rounds and subsequently completed eighty test trials. The outcome variable is reaction time (ms) of approach and avoidance for the appropriate stimuli. Following the task, participants returned to the original Qualtrics survey. They rated each encountered stimulus on three scales: Morality: From -100 (absolutely morally bad) to 100 (absolutely morally good), Affect: From -100 (absolutely aversive) to 100 (absolutely pleasant), and Complexity: From -100 (absolutely complex) to 100 (absolutely simple). After these ratings, participants were debriefed and concluded their participation.

**Analysis Approach**

To answer our first research question, we 1) used linear mixed models to evaluate if we can infer any relationships between variables in our data and moral approach-avoidance reaction times, 2) used supervised machine learning to examine the predictive relationship of potentially relevant features on response time, and 3) implemented unsupervised machine learning to identify patterns in the structure of the underlying data and provide a data driven validation of the stimuli used in this study. The data we used for all approaches contained 1) only accurate trials that were greater than or equal to 100ms and less than or equal to 2500ms and 2) came from participants with greater than or equal to 70% overall accuracy on all trials.

***Linear Mixed Effect Models (LMER)***

We aimed to identify the existence of moral approach-avoidance action tendencies by measuring whether individuals approach positive stimuli quicker (in comparison to approaching negative or neutral stimuli), and similarly whether individuals avoid negative stimuli quicker (in comparison to avoiding positive or neutral stimuli). To do this, we took a null-hypothesis testing approach to infer any existing relationships with reaction time.

We proceeded with the following hypotheses:

H1) People will exhibit implicit moral biases such that they will be quicker to approach morally good stimuli compared to (a) morally neutral (positive or negative valence) and (b) morally bad stimuli.

H2) they will be quicker to avoid morally bad stimuli compared to (a) morally neutral and (b) morally good stimuli.

H3) We expect the above effect to be larger with explicit evaluation than with implicit evaluation.

Since we are interested in confirming the existence of moral approach/avoidance actions and what factors lead to their manifestation, we are interested in establishing the association of possible factors with response time. Our data consists of several rounds of answers coming in from the same participant. Due to this, utilization of general linear models is limited as they assume that all data points are independent of each other, i.e. value of an observation coming from one participant does not affect the other observations. Therefore, we are adopting linear mixed effect models to account for the differences that we may see in response time based on the subject. We are interested in relating these differences in regard to both the biases that may be seen in relation to phrases and in relation to images, so two sets of models will be built. A series of three models were built for the two datasets, the first dataset with phrases as stimuli, and the second dataset with images as stimuli. The three models include one for implicit evaluation of the stimuli, one for the explicit evaluation of the stimuli, and one containing both implicit and explicit evaluation of the respective phrases or images. In the final combined approach, evaluation (implicit/explicit) type will be included as a binary indicator variable in our modeling process.

The model selection process involved iterative model building with inclusion/exclusion of possible factors affecting participant response time. Comparison of these models was done using Analysis of Variance (ANOVA). ANOVA allows us to compare model metrics such as model complexity, Bayesian Information Criterion (BIC), and model performance, Akaike Information Criterion (AIC). A preference to lower BIC was given over model performance as the goal of the study is inference rather than prediction.

***Supervised Machine Learning Regression Algorithms***

The goal, in line with our first research question of examining the existence of moral-approach-avoidance tendencies, was to quantify our ability to predict response time as well as inform us of which features in our data were offering the greatest predictive signal. This complements the null-hypothesis testing approach of examining if we can infer a relationship with response time. The approach laid out below was used for phrases and images aggregated across implicit and explicit evaluation as well as for both implicit and explicit evaluation separately.

The regression algorithms used included Ridge Regression (Ridge), Support Vector Regression (SVR), Random Forest (RF) Regressor, and XGBoost, whose implementations are provided by the Python machine learning library scikit-learn. The Ridge model was trialed using several solvers such as ‘svd’, ‘cholesky’, ‘lsqr’, and ‘sag’, coupled with several regularization strengths (alpha values from 1e-5 to 100). Additionally, configurations both inclusive and exclusive of an intercept term were explored. The SVR model was assessed by searching through multiple penalty parameter C and coefficient gamma values. Random Forest had maximum tree depth and minimum sample threshold hyperparameters searched through. Lastly, for XGBoost the number of boosting rounds, the tree’s maximum depth, and step size shrinkage were adjusted.

Model performance was evaluated by two distinct metrics, R-squared (R2) and Mean Absolute Error (MAE). R2 represents the proportion of the dependent variable’s variance that is attributable to the independent variables, or predictive features, with a possible perfect score of 1 (i.e. the further from 0, the better), while MAE is the mean value of the absolute differences between predicted and actual outcomes with the ideal MAE being minimized (i.e., closer to 0). It is important to note that as an absolute error metric, MAE is scale dependent. Thus, an interpretation of its magnitude should take into account the MAE's value relative to the range of the response variable.

Feature importance was deduced from the Random Forest Regressor model.The contribution of each feature in predicting response time was represented through bar plots. These plots show the mean decrease in node impurity for each feature as well as the associated standard deviations across trees.

***Unsupervised Machine Learning Algorithms***

In our exploratory data analysis, we were able to plot the distribution of the three categorical moral valences (Moral, Neutral, and Immoral) as they related to the parameters log response time and participant-rated rankings (Mortality, Affect, and Complexity). Utilizing unsupervised machine learning, we sought to recreate the distribution patterns we created in our exploratory data analysis. We utilized a clustering approach with k = 3 to categorize responses into morally good, neutral, and morally bad clusters. If these categories accurately mirrored the results from plotting based on labels, we would be able to progress with attempting to predict moral valence based on other parameters, including participant ratings. K-means was the first model utilized because of its simplicity.

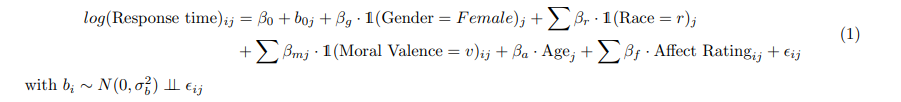
## **Results**

**Linear Mixed Effects Models (LMER)**

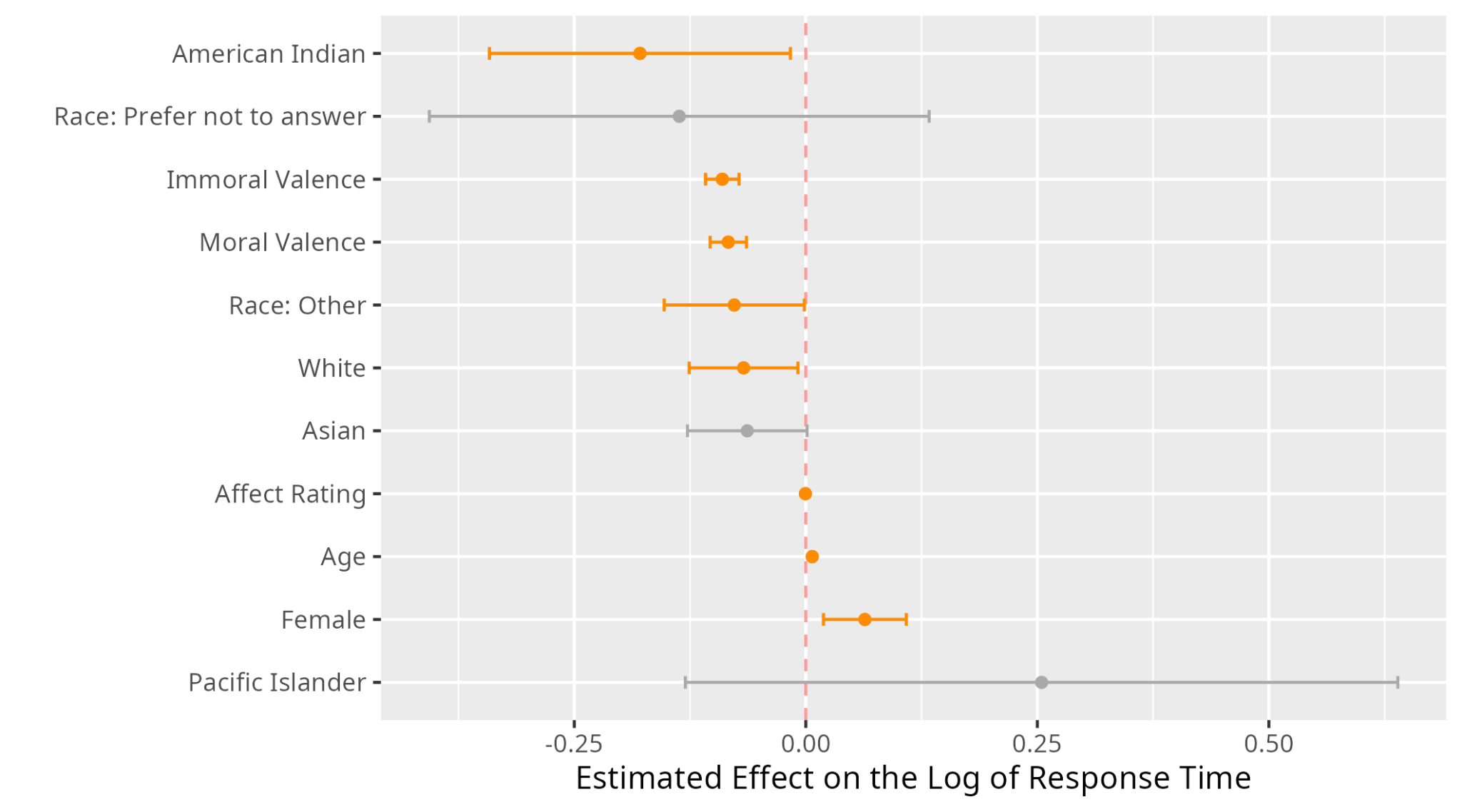
***Phrases:*** *Explicit*

In the first set of models, we looked at the moral approach-avoidance task where participants respond to phrases that have varying moral valence (i.e., morally good, bad, or neutral). The first model we look at exclusively focuses on explicit biases, that is that individuals were told to approach morally good stimuli and avoid morally bad stimuli.

The final model to predict the (log) response time based on the ***Explicit-Phrases*** data is as follows:



We can observe the variable influence to the log of response time in the following plot. Table A1 (Appendix) contains the exact point and interval estimates.



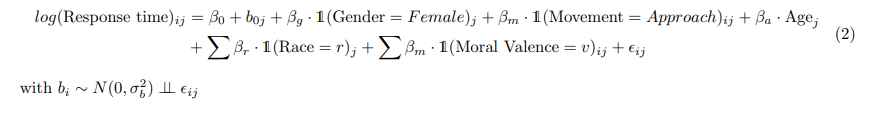
The figure above (and all similarly constructed plots below) shows point estimates and 95% confidence bands for the model parameters. The intervals highlighted in orange are significant (do not cross the dashed red line at 0), and the intervals that are in gray do cross the dashed red line. For categorical variables such as gender or moral valence,the red-dashed line refers to where the baseline category would lie. For example, since the 95% confidence band for Female is strictly above the dashed red line, we conclude that females have a slower reaction time than males. The point estimate for Female is 0.06 (Table 1A) with the gender male as reference, which suggests that we would expect to see a response time that is 0.06 log-milliseconds slower.

Considering this same example, we perhaps will be more interested in looking at milliseconds, and not log-milliseconds. By exponentiating the parameter estimates, we can look at the multiplicative change in the response when compared to the baseline category for a categorical variable. We would expect the response time to have a multiplicative change of 1.06 if the participant is female than if male, controlling for all other variables.

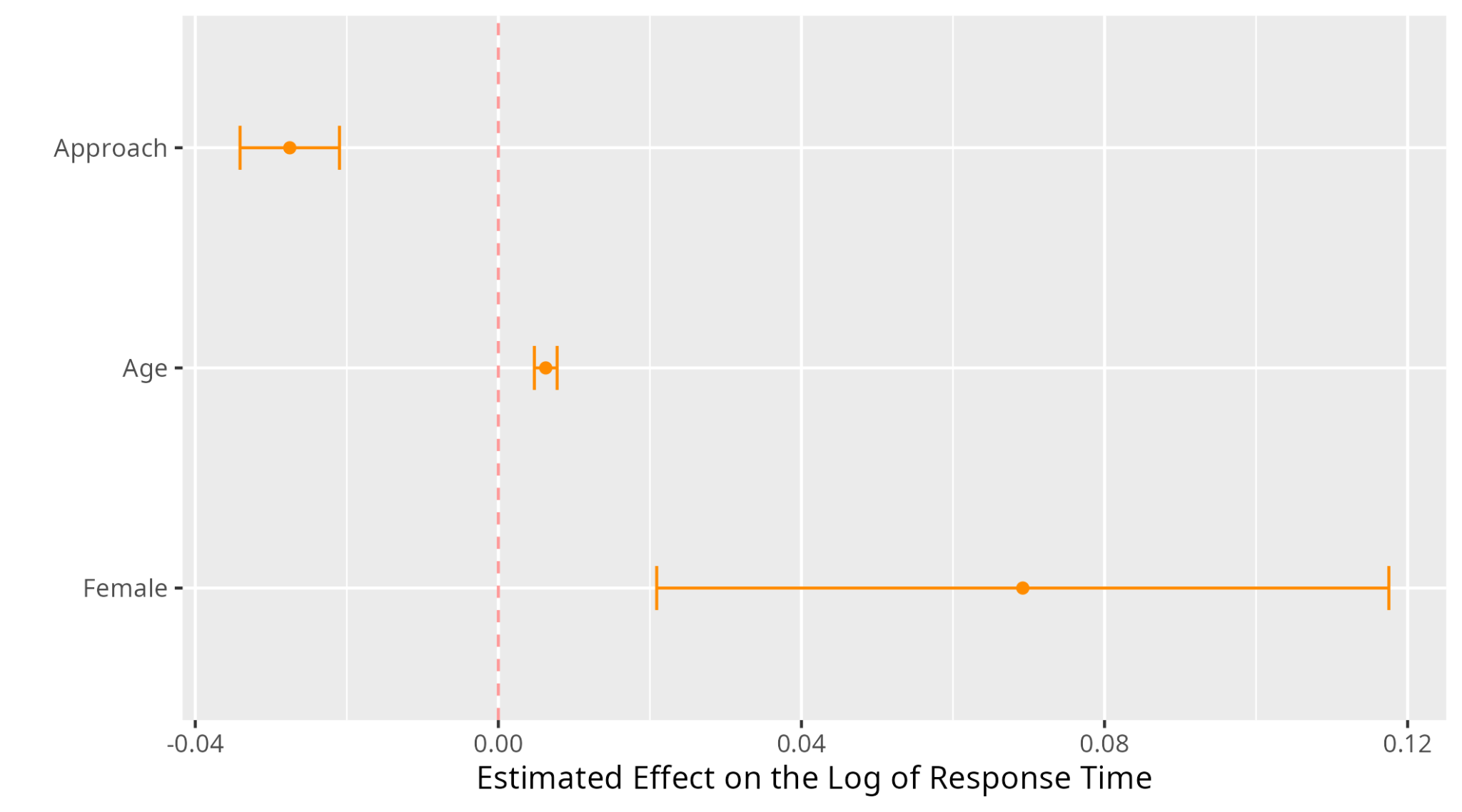
Notable interpretations for the Explicit-Phrases model include that after controlling for all other variables, the response time is estimated to increase by 6.6% if the participant is a female. Immoral valence suggests a decrease in the reaction time by 8.61% when compared with neutral morality; moral valence suggests a decrease in the reaction time by 7.69%. Each point increase in affect rating (0 = unpleasant, 100 = pleasant) is estimated to decrease response time by 1 ms. With each additional year in age, the estimated response time will increase by 1 ms.

***Phrases:*** *Implicit*

Secondly, we looked at phrases with implicit biases, which is when the individual is told to approach or avoid based on the color of the background to the phrase, regardless of the moral content.

The final model based on the ***Implicit-Phrases*** data is as follows:

The figure below highlights all of the significant variable’s influence in terms of predicting the log of response time. Table A2 contains the exact point and interval estimates.

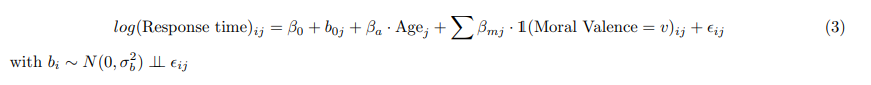


For the Implicit-Phrases dataset, our model suggests that females have a slower reaction time than males, with a 1-exp(0.069) = 7.14% increase compared to males, controlling for all other variables. Each additional year in age will see reaction time increase by 1 ms. Approaching is faster than avoiding, decreasing the reaction time by 2.66%. (NB: the estimates on the graph are log-response time. Estimates have been exponentiated to measure response time).

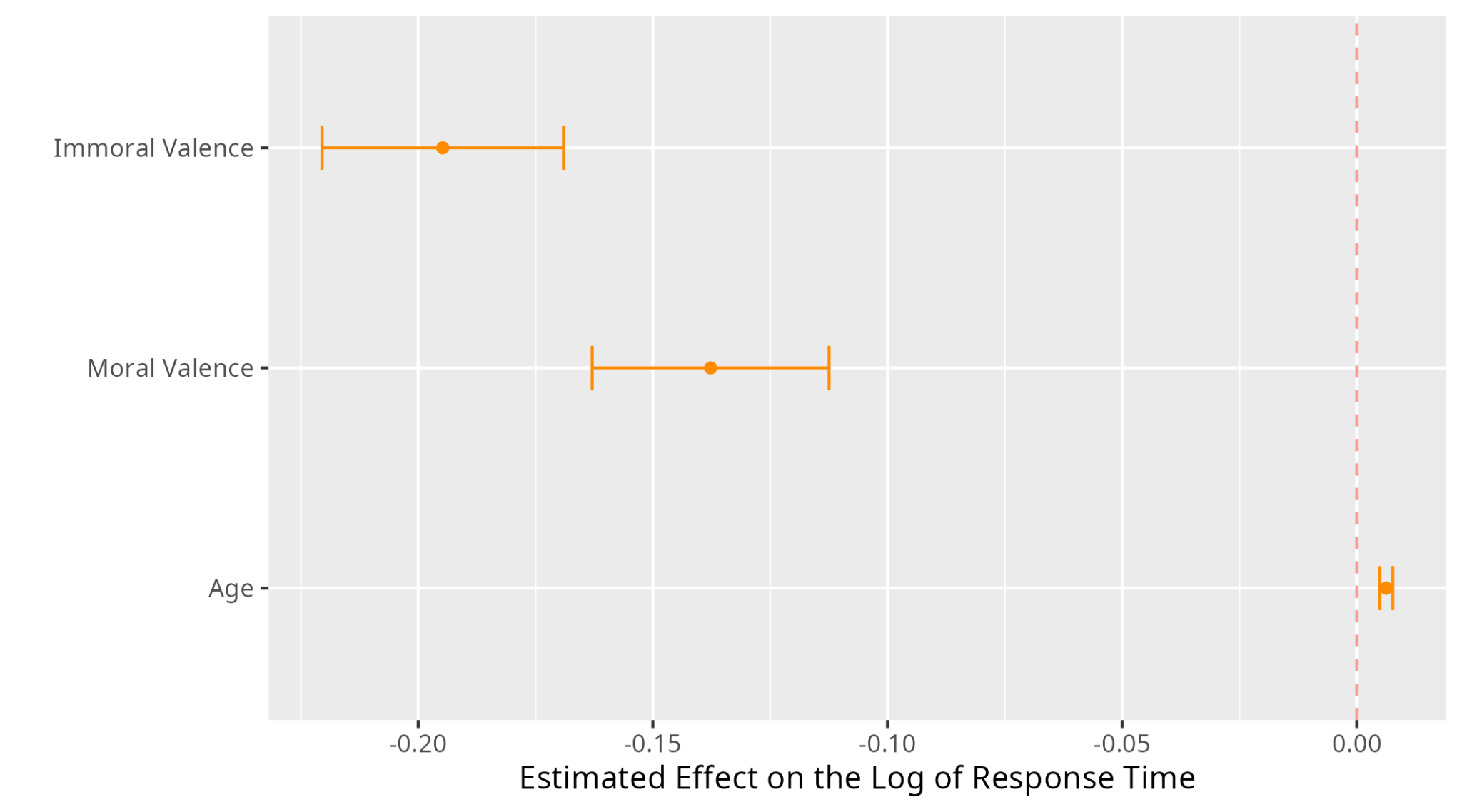
**Images:** *Explicit*

We then repeated the same investigations in the images dataset, where participants responded to images with varying moral valence, which we presume to be different from just phrases as images can provoke a more heightened emotional response compared to phrases alone. We did a similarly structured analysis, where we exclusively model data with either explicit or implicit bias.

The final model based on the ***Explicit-Images*** data is as follows,



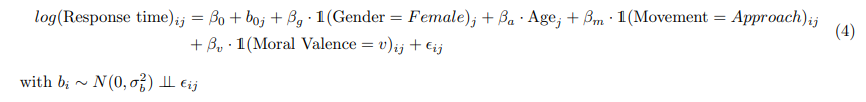
and the following plot highlights the impact of the variables in the model. Table A3 contains the exact point and interval estimates.



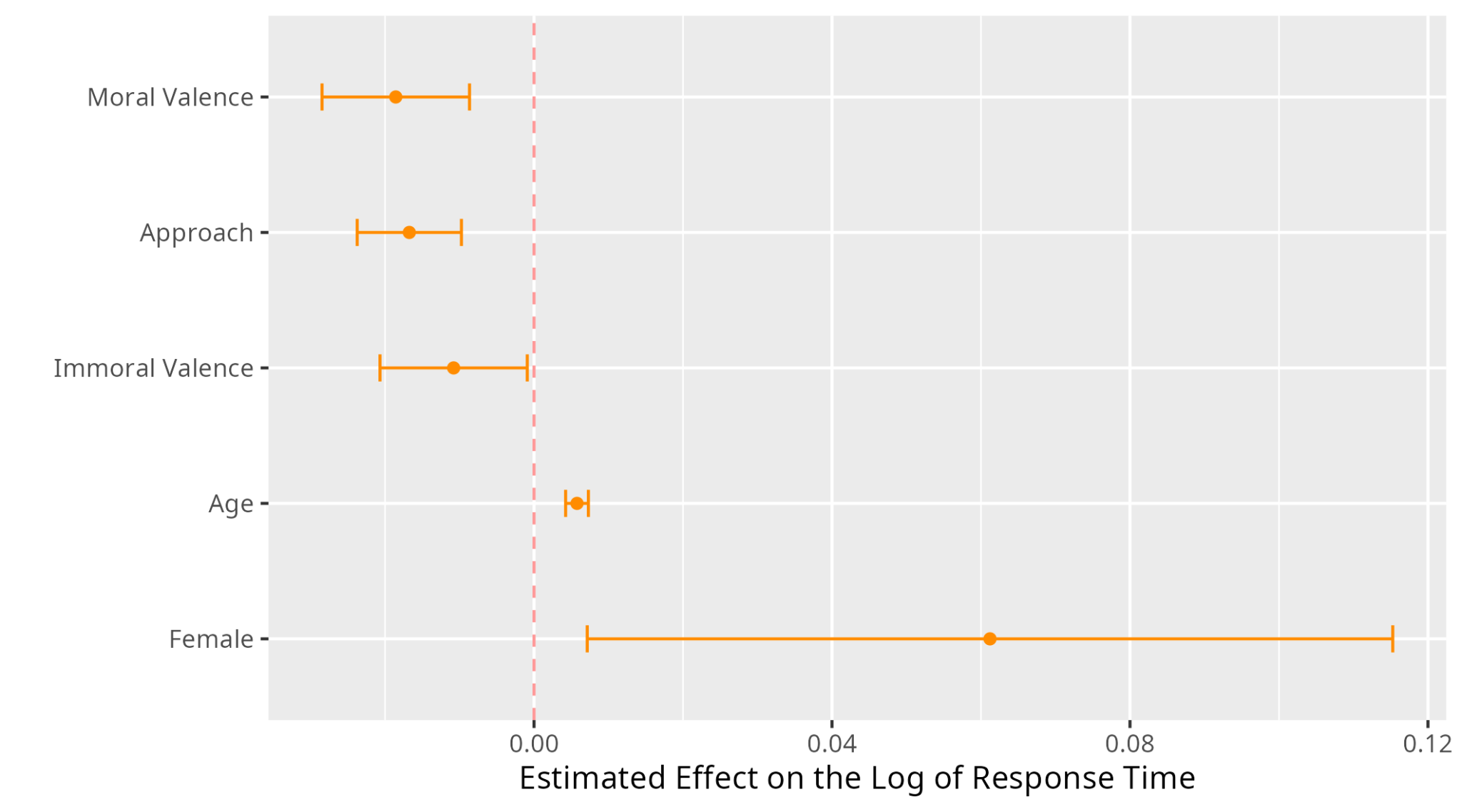
For the Explicit-Images, our model suggests that after controlling for all other variables, each additional year of age suggests an increase in response time by 1 ms. Response time is estimated to decrease if the valence of the image is immoral (compared to neutral) by 17.71%. The response time is also estimated to decrease if the image valence is moral (compared to neutral), by 12.89%.

***Images:*** *Implicit*

Similarly, we built a model for the ***Implicit-Images*** data, where the final model is as follows,



The following plot highlights the impact of the variables in the model. Table A4 contains the exact point and interval estimates.



For the Implicit-Images model, to summarize notable coefficient estimates, when controlling for all other variables, each additional year in age is estimated to slow response time by 1 ms. If the participant is female, we also expect her reaction time to be 6.18% slower than a male, keeping all else constant. Approaching is significantly faster than avoiding, with an estimated decrease in response time of 1.66%. Response time is estimated to decrease if the valence of the image is immoral (compared to neutral) by 1.09%The response time is also estimated to decrease if the image valence is moral (compared to neutral), by 1.

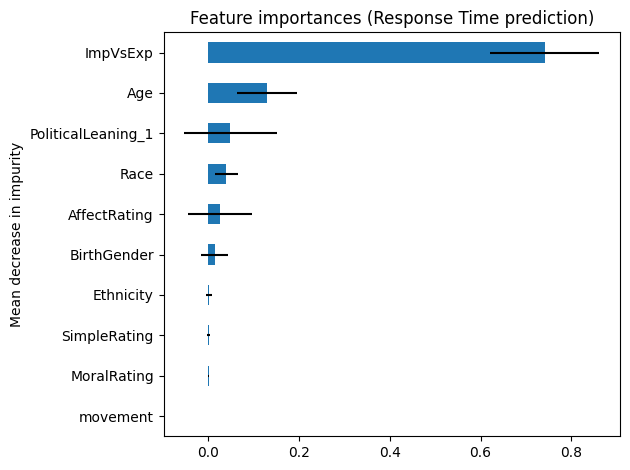
**Summary**

Overall, we noticed that females generally had slower reaction times compared to males (and consistently showed this trend in all of the models). We may be interested in performing a literature review to see why that may be the case. It is less surprising, however, that reaction time would increase with age, as physical and mental sharpness declines. Neutral moral valence consistently showed the slowest response time. In the implicit data models, moral valence was the fastest (but moral valence was only significant in the images model); in explicit models, immoral valence has faster response time. Approaching versus avoiding was only significant in the implicit models, where approaching was seen to be faster than avoiding.

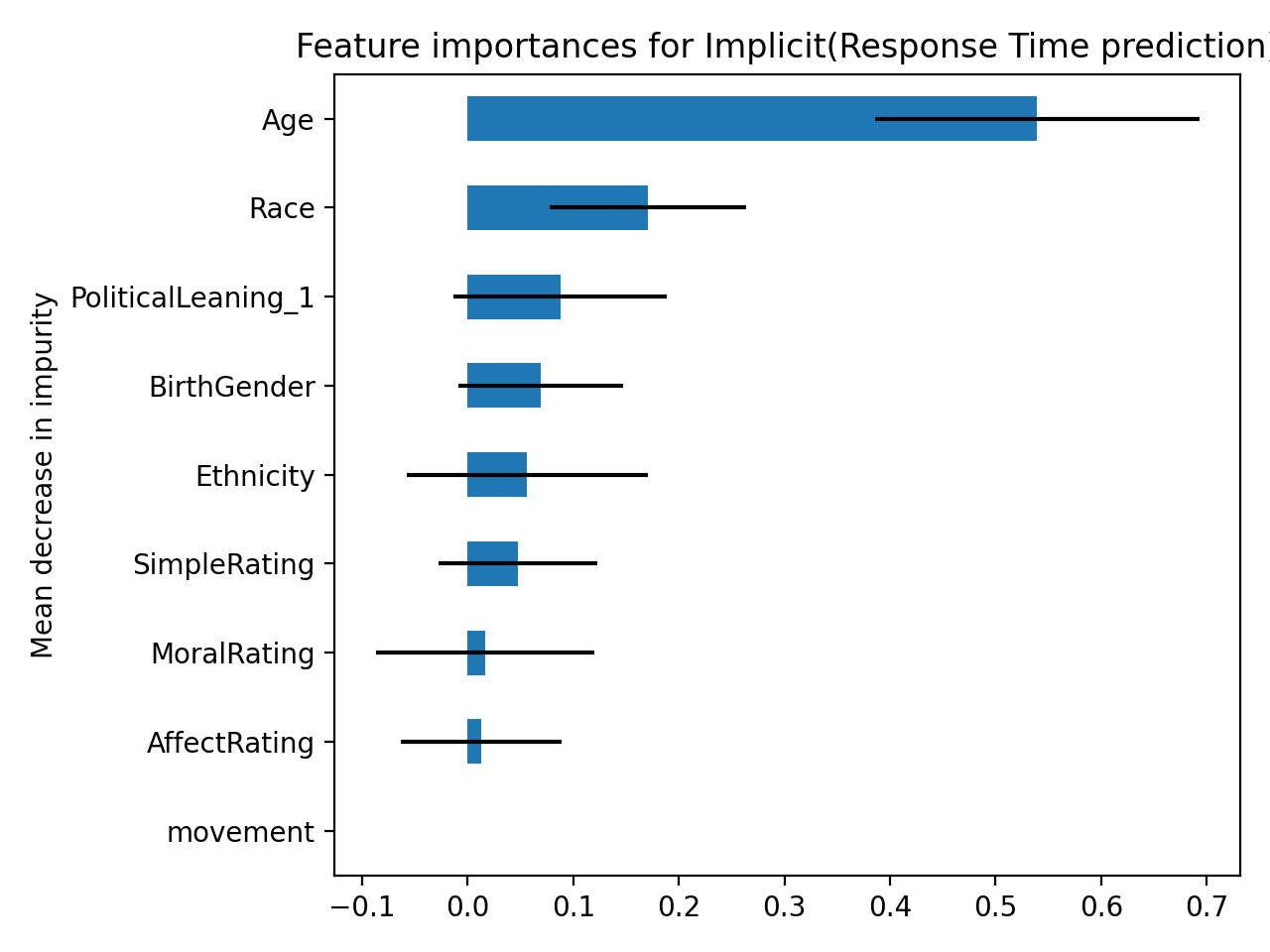
We also created two models (one for phrases, one for images), where whether or not the prompt was to respond to explicit vs implicit biases was included as a covariate in the model, rather than dividing into two separate datasets. Under the full model, we noticed that contrary to our hypothesis, implicit bias was much quicker to respond to than explicit bias. This could be because individuals are actively not looking at the content at hand, but that the ‘color of the background’ is simply obvious and doesn’t take much processing time. If that is the case, this is a flaw in the design of the experiment, and so results would need to be taken with this potential influence in mind.

**Supervised Machine Learning Regression Algorithms**

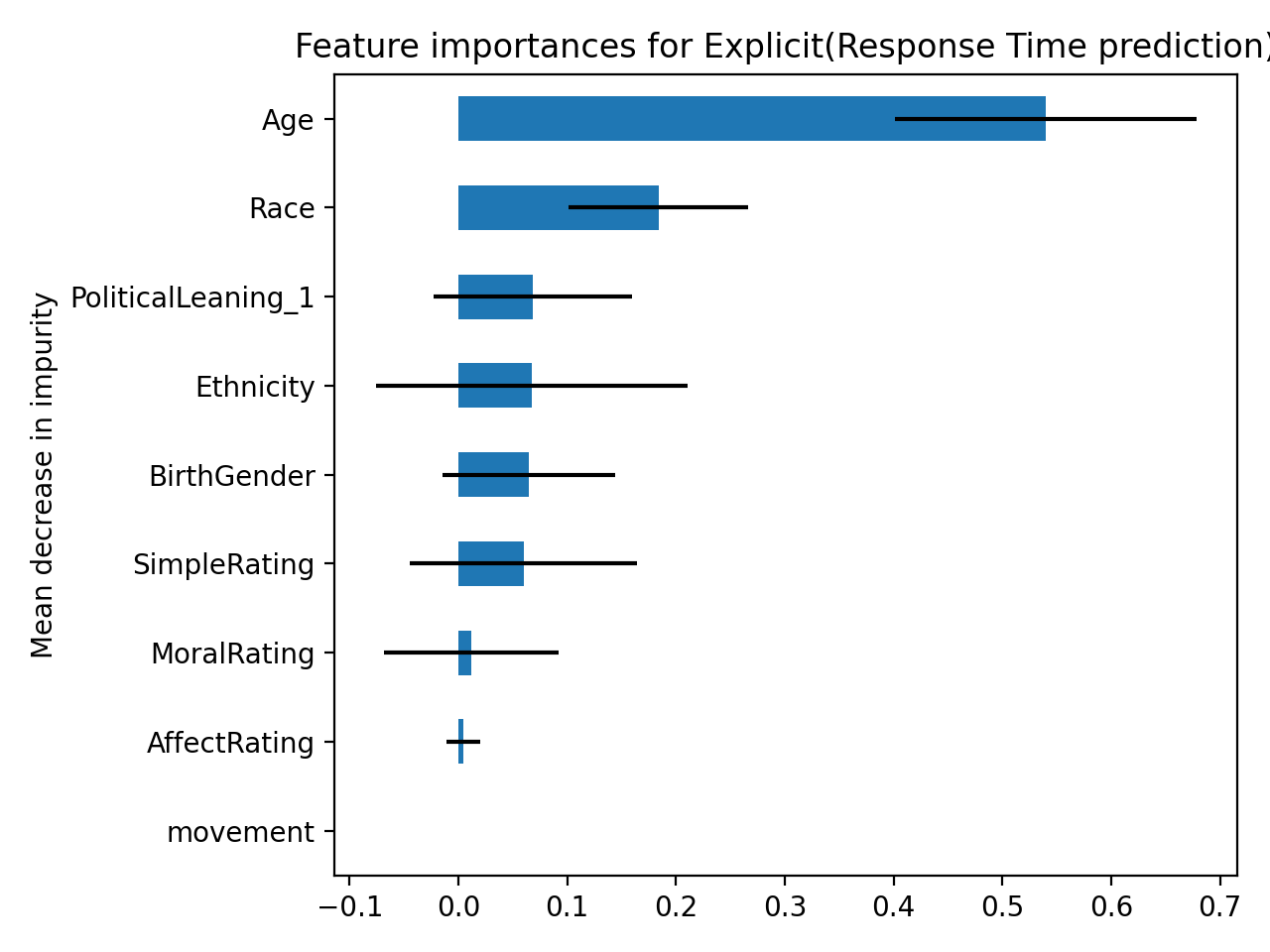
***Phrases***

**Full Data.** Among the models, SVR performed the best with an R2 value of 0.143 and an MAE of 292.59 for predicting response time. The rest of the models did not perform very well and had R2 values from -0.034 to 0.072. Additionally, the MAE for these models were higher up to a maximum of 383.15. The features that carried the most weight were type of evaluation (ImpVsExp) and Age. By far the type of evaluation had the lion’s share of the predictive power with close to .7 mean decrease in impurity. This is inline with the differences show between implicit and explicit evaluation with the linear mixed models above. Everything else was below 0.10. Moral Rating did not reliably decrease impurity. Overall, the ability to predict response time is not very high for aggregate phrase data.

|  | **R2** | **MAE** |
| --- | --- | --- |
| **Ridge** | 0.069 | 383.15 |
| **RF** | 0.072 | 340.67 |
| **SVR** | 0.143 | 292.59 |
| **XGBoost** | -0.034 | 335.63 |

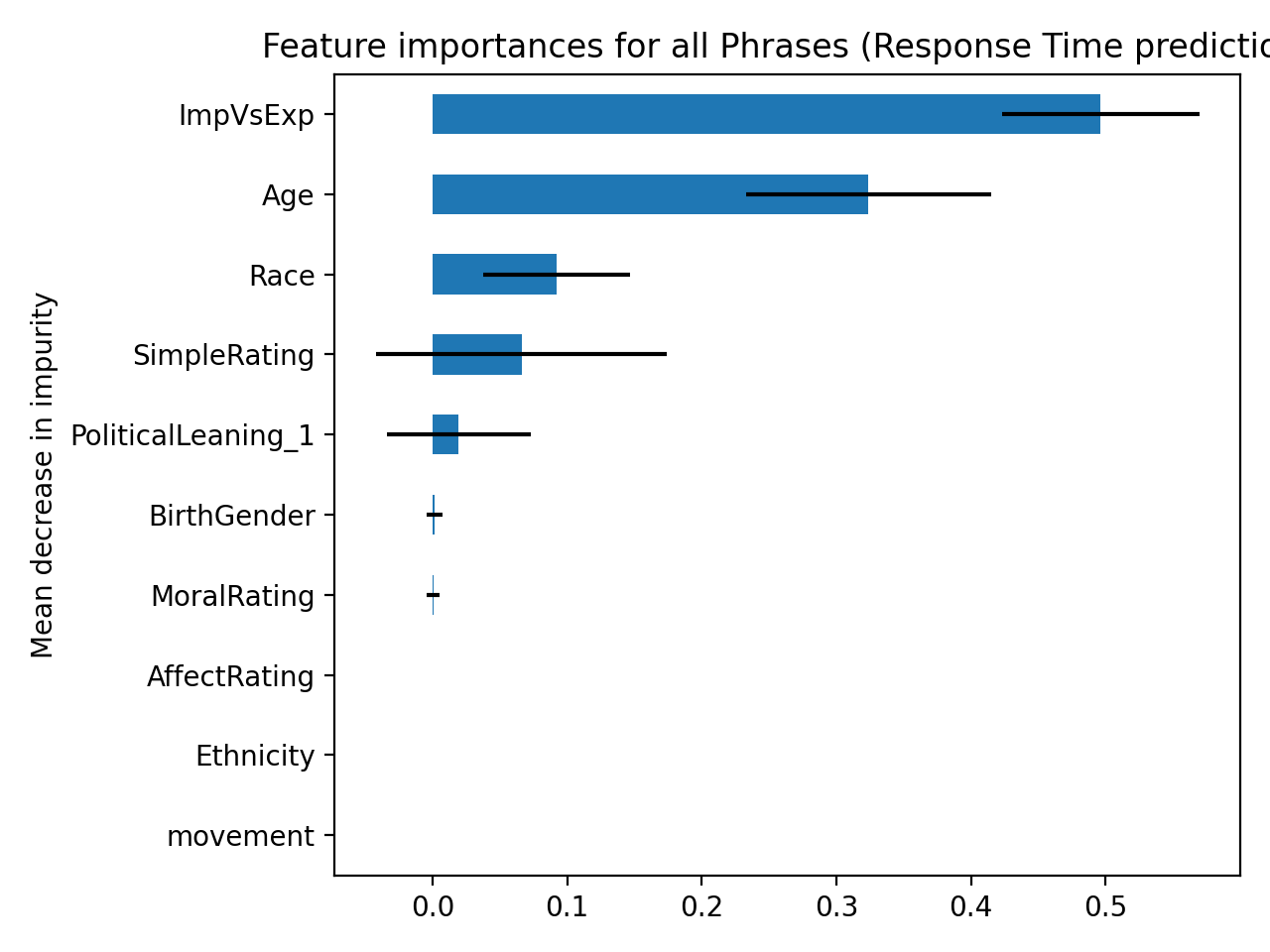
**Implicit.** Among the models, SVR performed the best with an R2 value of 0.112 and an MAE of 311.55 for predicting response time. This decrease in predictive ability is inline with a lack of meaningful differences in implicit response times in the LMERs. The other models had poor performance (i.e., R2 close to “0”). Additionally, the MAE for these models were high up to a maximum of 439.51. The features that carried the most weight were Age and Race. Age had the lion’s share of the predictive power with close to .55 mean decrease in impurity. Everything else was below 0.10. Moral Rating did not reliably decrease impurity. Overall, the ability to predict response time is poor for implicit phrase data.

|  | **R2** | **MAE** |
| --- | --- | --- |
| **Ridge** | 0.008 | 439.51 |
| **RF** | 0.020 | 374.40 |
| **SVR** | 0.112 | 311.55 |
| **XGBoost** | 0.036 | 343.68 |

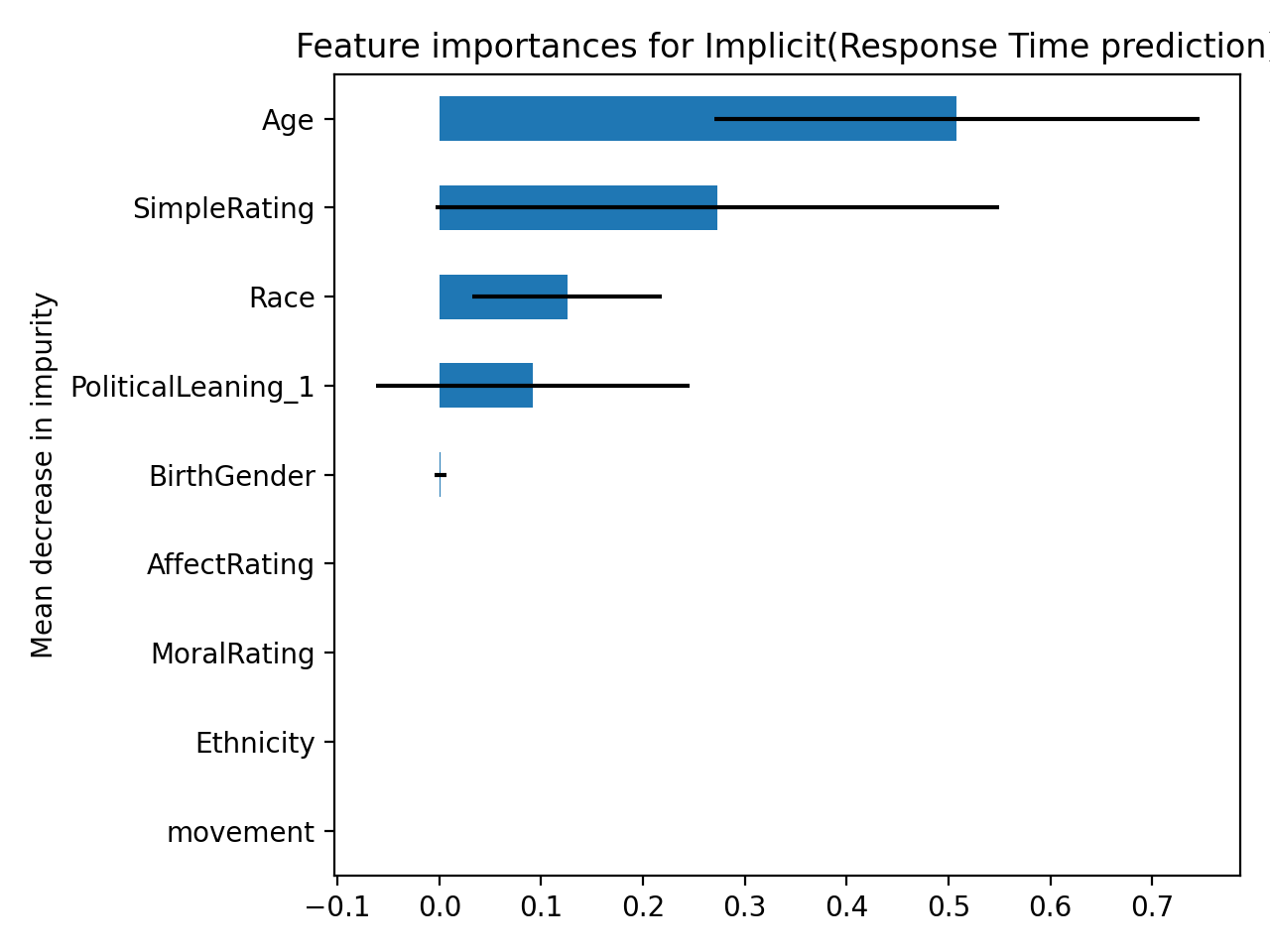
**Explicit.** Among the models, SVR performed the best with an R2 value of 0.145 and an MAE of 304.22 for predicting response time. This, along with the rest of models, is essentially floor performance (i.e., R2 = “0”). Additionally, the MAE for these models were high up to a maximum of 429.23. Again, the features that carried the most weight were Age and Race. Age had the lion’s share of the predictive power with close to .55 mean decrease in impurity. Everything else was below 0.10. Moral Rating did not reliably decrease impurity. Overall, the ability to predict response time is maybe marginally better for than implicit but still relatively poor for explicit phrase data.

|  | **R2** | **MAE** |
| --- | --- | --- |
| **Ridge** | 0.012 | 429.2 |
| **RF** | 0.022 | 361.56 |
| **SVR** | 0.145 | 304.22 |
| **XGBoost** | 0.044 | 344.32 |

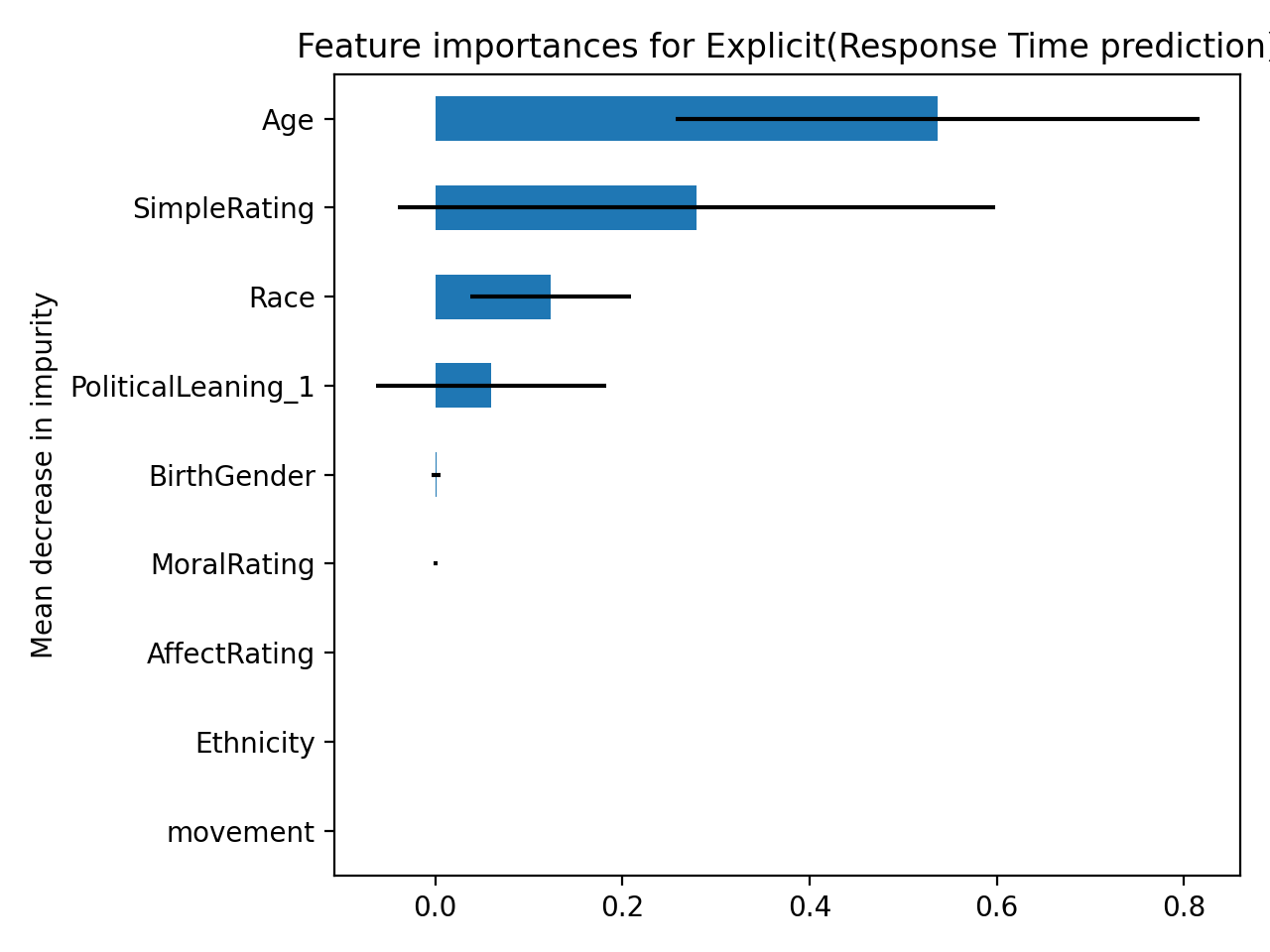
***Images***

**Full Data.** Among the models, RF performed the best with an R2value of 0.029 and an MAE of 362.24 for predicting response time. This, along with the rest of models, is essentially floor performance (i.e., “0”). Additionally, the MAE for these models were high up to a maximum of 396.68. Similar to image aggregate, the features that carried the most weight were type of evaluation (ImpVsExp) and Age. Although to a lesser extent than with phrase data, type of evaluation had the majority of the predictive power with close to .5 mean decrease in impurity. Age had close to 0.35. Everything else was below 0.10. Moral Rating did not reliably decrease impurity. Overall, the ability to predict response time is very poor for aggregate image data.

|  | **R2** | **MAE** |
| --- | --- | --- |
| **Ridge** | 0.027 | 396.68 |
| **RF** | 0.029 | 364.24 |
| **SVR** | -0.009 | 371.82 |
| **XGBoost** | -0.044 | 358.82 |

**Implicit.** Among the models, Ridge performed the best with an R2 value of 0.009 but with an MAE of 414.99 for predicting response time. This, along with the rest of models, is essentially floor performance (i.e., “0”). The features that carried the most weight were Age, Simple Rating, and Race. Age had the lion’s share of the predictive power with close to 0.50 mean decrease in impurity. Everything else was at or below 0.10. Moral Rating did not reliably decrease impurity. Overall, the ability to predict response time is very poor for implicit image data.

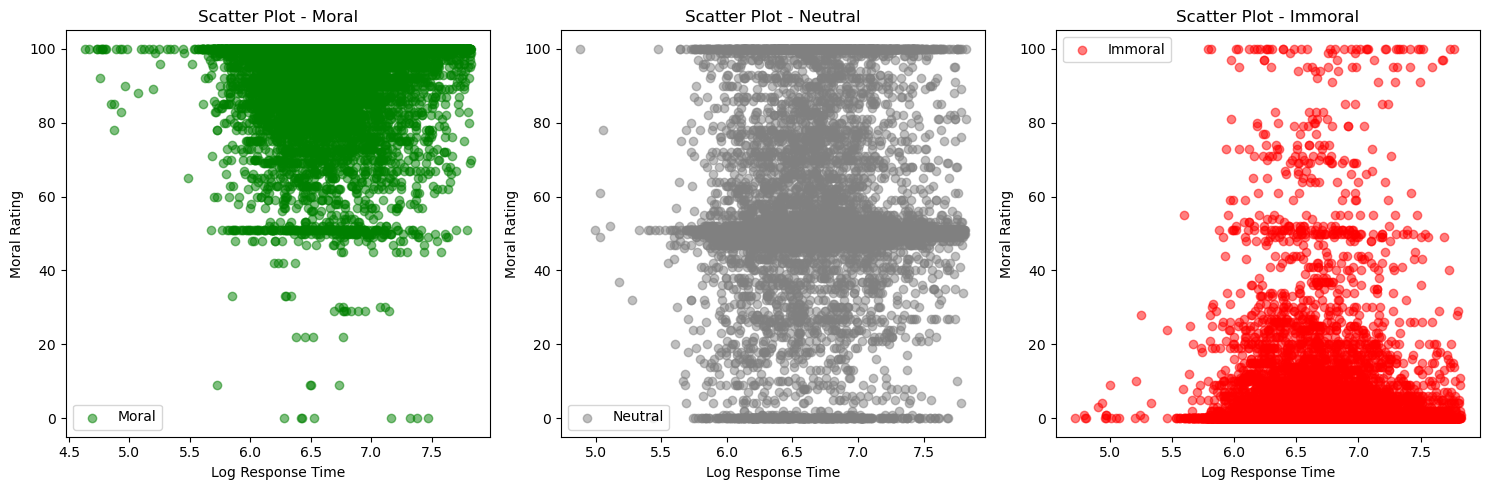
|  | **R2** | **MAE** |
| --- | --- | --- |
| **Ridge** | 0.009 | 414.99 |
| **RF** | -0.001 | 378.43 |
| **SVR** | -0.011 | 374.95 |
| **XGBoost** | -0.084 | 362.91 |

**Explicit.** Among the models, Ridge performed the best with an R2 value of 0.009 but with an MAE of 414.99 for predicting response time. This, along with the rest of models, is essentially floor performance (i.e., “0”). The features that carried the most weight were Age, Simple Rating, and Race. Age had the lion’s share of the predictive power with close to .60 mean decrease in impurity. Everything else was at or below 0.10. Moral Rating did not reliably decrease impurity. Overall, the ability to predict response time is very poor for explicit image data. It is still unclear if there is any meaningful signal or consequential interpretation to be made about the reason SVR was the best model in some cases but in other modeling attempts it was Ridge, for example.

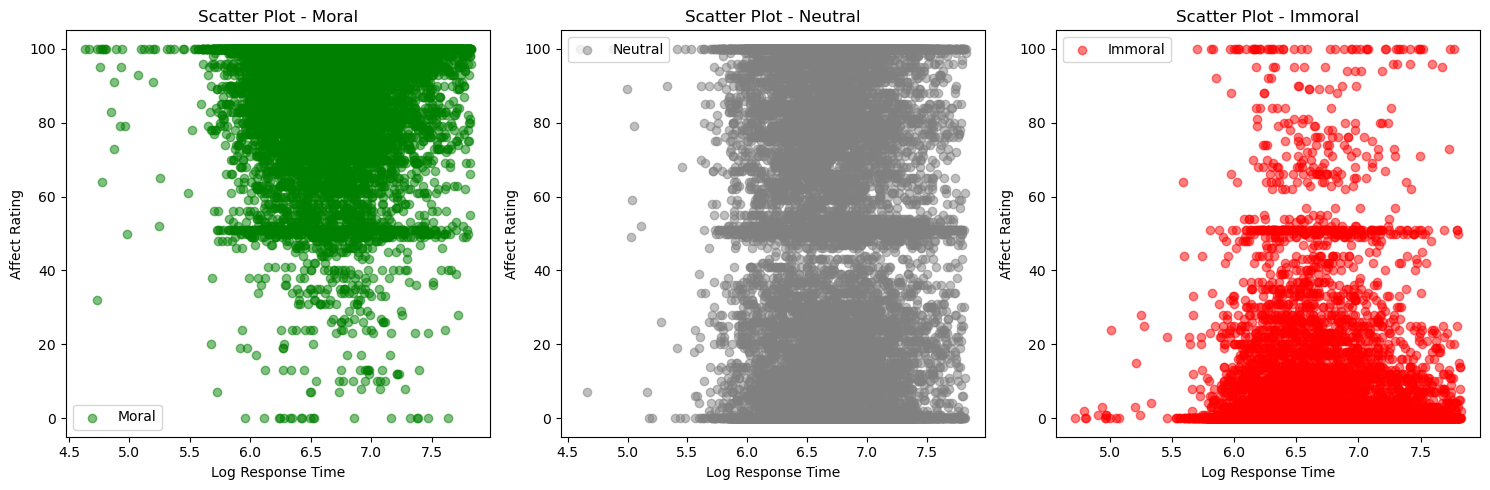
|  | **R2** | **MAE** |
| --- | --- | --- |
| **Ridge** | 0.009 | 414.99 |
| **RF** | -0.001 | 378.09 |
| **SVR** | -0.011 | 374.94 |
| **XGBoost** | -0.084 | 362.91 |

**Unsupervised Machine Learning Algorithms**

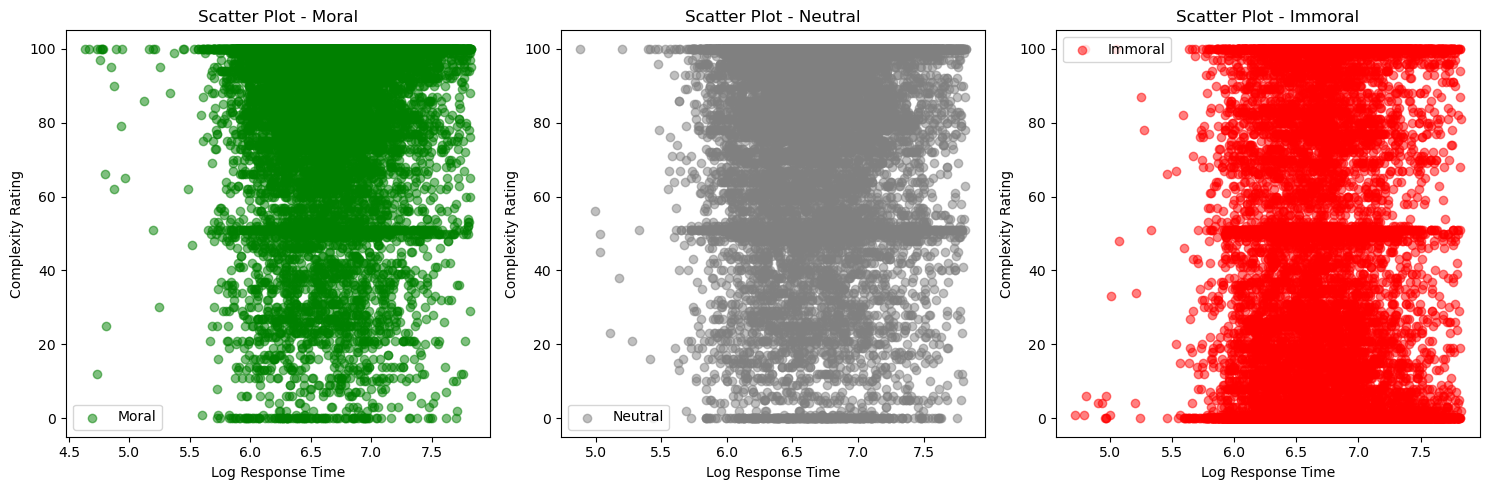
The exploratory analysis of how different moral valences (Moral, Neutral, Immoral) relate to log response time and participant-rated categories of Morality, Affect, and Complexity are shown below. Figures depicting Moral ratings by participants reflected expected results, with actions labelled as Moral being most frequently judged as moral by participants. Likewise, actions labelled as Neutral were concentrated at an average level with equal distribution on both sides, and actions labelled as Immoral were labelled as low morality by participants.



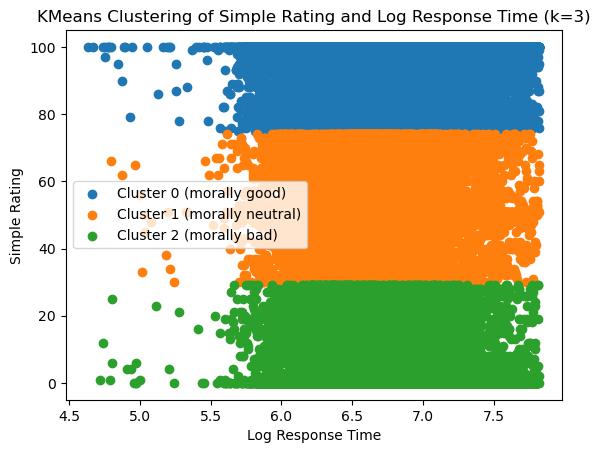
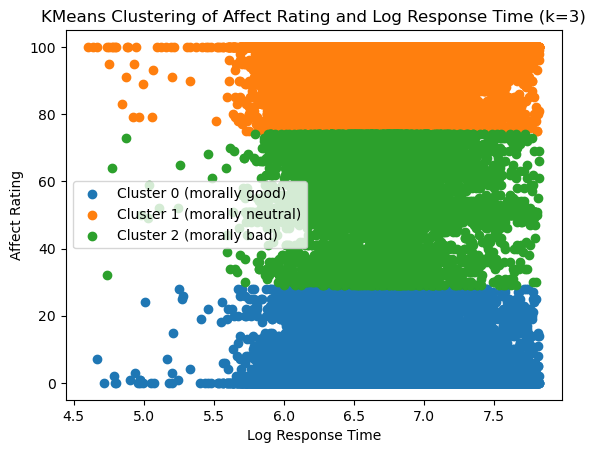
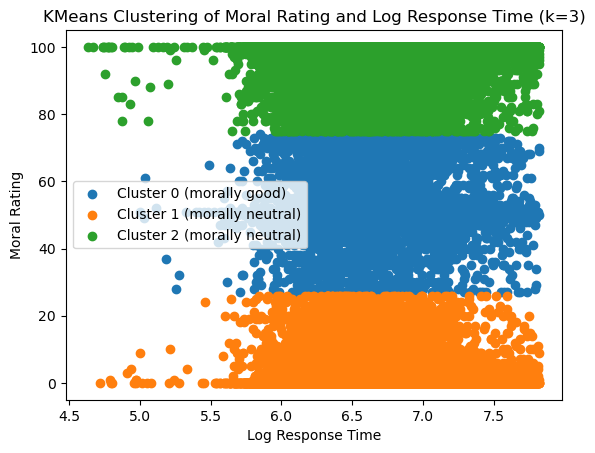
Actions labelled as morally good were rated as having high affect by participants, indicating participants view these actions as pleasant or positive. Neutral actions were rated evenly across the spectrum at different affect, and actions labelled Immoral were rated as unpleasant or negative.



Actions were rated evenly with a skew toward less complex participant ratings. Immoral actions were evenly distributed across complexity ratings.



The K-Means clustering analysis was applied to examine the relationships between moral valences, affect, and simplicity ratings in relation to log response time, utilizing a kvalue of 3 to represent the categorical valences of morally good, neutral, and morally bad. The results revealed a consistent and similar spread of data, with the majority of data points demonstrating log response times falling within the range of 5.5 to 7.5. This was similar to the pattern found in exploratory data analysis.



However, the distribution of data points across the three clusters, each corresponding to the categorical valences, do not align with the labelled groups established by EDA. This demonstrates that K-Means is not an effective model to achieve our goal of predicting labels based on participant ratings of morality, affect, and complexity. Moving forward, different iterations on K-means and different methods of unsupervised learning will be used to better approximate the distribution of data. Additionally, different unsupervised learning algorithms that focus on categorical variables will be utilized on applicable variables in the dataset.

## **Future Directions**

In advancing our research, it’s important to recognize that even with the large amount of work that we have already accomplished, there is still much to be done. It is imperative to focus on both optimizing and expanding the focus of our machine learning approaches. With our supervised models, there is a need to refine our models for not just predicting response time, but also to predict the moral rating given to each stimulus as well as predicting the specific stimulus reacted to and accurately categorizing these stimuli into moral valence categories. Regarding our unsupervised approaches, we will introduce more sophisticated models to enhance the clustering of data. Additionally, we will run our unsupervised approaches not just on individual observations as we have already done, but also for stimuli and participants to explore inherent similarities and patterns.

Additionally, incorporating a simulation-based power analysis will simulate various scenarios to assess the study’s ability to detect true effects. This will help ensure the robustness of our findings and guide future study designs and participant recruitment. Furthermore, exploring the asymmetries and differences between the inferential and predictive relationships between our data is essential. Lastly, comparing the moral valence assigned by participants versus that pre-determined by us will offer valuable perspectives on the subjective experience or moral approach-avoidance tendencies.

## 

## 

## **Resources**

Aubé, B., Rougier, M., Muller, D., Ric, F., & Yzerbyt, V. (2019). The online-VAAST: A short and online tool to measure spontaneous approach and avoidance tendencies. *Acta Psychologica*, *201*, 102942.<https://doi.org/10.1016/j.actpsy.2019.102942>

Cameron CD, Payne BK, Sinnott-Armstrong W, Scheffer JA, Inzlicht M. Implicit moral evaluations: A multinomial modeling approach. Cognition. 2017 Jan;158:224-241. doi: 10.1016/j.cognition.2016.10.013. Epub 2016 Nov 16. Erratum in: Cognition. 2018 Jan 4;: PMID: 27865113.

Strohminger, Nina & Caldwell, Brendan & Cameron, Daryl & Schaich Borg, Jana & Sinnott-Armstrong, Walter. (2014). Implicit Morality: A Methodological Survey. 10.1057/9781137409805\_10.

## **Appendix**

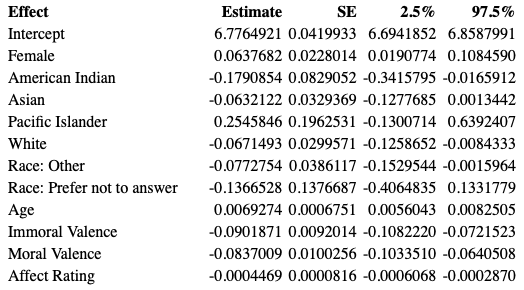


Table A1: Point Estimates and Confidence Intervals for Explicit-Phrases

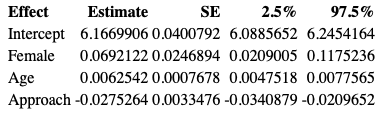


Table A2: Point Estimates and Confidence Intervals for Implicit-Phrases

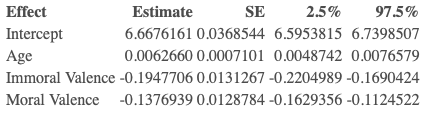


Table A3: Point Estimates and Confidence Intervals for Explicit-Images

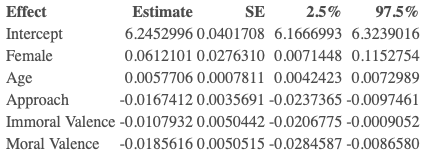


Table A4: Point Estimates and Confidence Intervals for Implicit-Images