The good person is me: Spontaneous self-referential process prioritizes moral character in perceptual matching

Hu Chuan-Peng[1, 2], Kaiping Peng[2], & Jie Sui[3]

[1] Nanjing Normal University, 210024 Nanjing, China

[2] Tsinghua University, 100084 Beijing, China

[3] University of Aberdeen, Aberdeen, Scotland

Author Note

Hu Chuan-Peng, School of Psychology, Nanjing Normal University, 210024 Nanjing, China. Kaiping Peng, Department of Psychology, Tsinghua University, 100084 Beijing, China. Jie Sui, School of Psychology, University of Aberdeen, Aberdeen, Scotland. Authors contriubtion: HCP, JS, & KP design the study, HCP collected the data, HCP analyzed the data and drafted the manuscript. All authors read and agreed upon the current version of the manuscripts.

Correspondence concerning this article should be addressed to Hu Chuan-Peng, School of Psychology, Nanjing Normal University, Ninghai Road 122, Gulou District, 210024 Nanjing, China. E-mail: hcp4715@hotmail.com

Abstract

Moral character is central to social evaluation and moral judgment. However, whether moral character information is prioritized in perceptual decision-making was debated. Here we investigated the effect of moral character on perceptual decision-making through an associative learning task. Participants first learned associations between different geometric shapes and moral characters and then performed a simple perceptual matching task. Across five experiments ($N = 192$), we found a robust prioritization effect of good character-related information, i.e., participants responded faster and more accurately to shapes that were associated with good characters than shapes associated with neutral or bad characters. We then examine whether the prioritization of good character was due to valence alone or an interaction between valence and self-reference. Data from three experiments ($N = 108$) demonstrated that the prioritization effect of good character was robust when the good character referred to the self but weak or non-exist when it referred to others. Additional two experiments ($N = 104$) further revealed that the mutual facilitation between good character and self-reference occurred even when one of them was task-irrelevant. Together, these results suggested a spontaneous self-referential process as a mechanism of the prioritization effect of good character.

Keywords: Perceptual decision-making, Self positivity bias, moral character

Word count: X

The good person is me: Spontaneous self-referential process prioritizes moral character in perceptual matching

Introduction

Is moral information prioritized in perception? This question has evoked much heat a few years ago but remains unsolved. On the one hand, morality is a basic dimension in social evaluation (Dunbar, 2004; Ellemers, 2018; Goodwin, 2015; Goodwin, Piazza, & Rozin, 2014), this importance should grant moral information more salient than morally neutral information and thus prioritized when the attentional resource is limited. This logic is similar to other stimuli that are also important to humans, e.g., threatening stimuli (e.g., Ohman, Lundqvist, & Esteves, 2001), rewards (B. A. Anderson, Laurent, & Yantis, 2011), or self-related stimuli (Sui & Rotshtein, 2019). Indeed, previous studies reported bad characters are prioritized in visual processing (E. Anderson, Siegel, Bliss-Moreau, & Barrett, 2011; Eiserbeck & Abdel Rahman, 2020), suggesting that bad people are detected faster than neutral or good people. On the other hand, there is evidence against the view that morally bad information is prioritized in perception. First, researchers reported positive bias in processing moral-related information. For example, Shore and Heerey (2013) found that faces with positive interaction in a trust game were prioritized in the pre-attentive process. Second, the negative bias in perceiving moral information is not robust (Stein, Grubb, Bertrand, Suh, & Verosky, 2017). Third, the mechanism underlying the reported negative bias in processing moral-related information is debated (Firestone & Scholl, 2015, 2016b; Jussim, Crawford, Anglin, Stevens, & Duarte, 2016). In short, while the importance of morality is widely recognized, whether moral information is prioritized in perceptual decision-making is still an open question. Here we manipulated the moral character by an associated learning task and investigated whether immediately acquired moral character information is prioritized in a perceptual matching task.

If moral character information is indeed prioritized, the next question is how?

Previous studies explain the effect based on valence. For example, the negative bias toward moral information is explained by aligning moral information with affective stimuli and threat detection was supposed to be the potential mechanism (B. A. Anderson et al., 2011). The positive bias toward moral information, on the other hand, is explained by value-based attention (Shore & Heerey, 2013). However, these explanations often ignore the fact the value is subjective *per se* (Juechems & Summerfield, 2019). Merely associating with the self can prioritize the stimuli in perception, attention, working memory, and long-term memory (Sui & Humphreys, 2015; Sui & Rotshtein, 2019). Here, we explicitly included self-relevance in our experimental design and tested whether the prioritization of moral character is modulated by self-relevance. We adopted an associative learning task, or self-tagging task, which has been widely used in studying the self-relevance effect. It is based on the well-established fact that humans can quickly learn the associations between symbols via language and change subsequent behaviors accordingly. This associative learning is widely used in aversive learning and value-based learning (Atlas et al., 2022; Deltomme, Mertens, Tibboel, & Braem, 2018). By explicitly instructing participants on which moral character is self-referencing and which is not, we can test whether the prioritization of moral character is by valence *per se* or by the self-referential of moral valence.

We address these questions by investigating how immediately acquired moral character information modulates the processing of neutral geometric shapes in a perceptual matching task. Unlike previous studies relies on faces or words as materials, stimuli used in the social associative task are geometric shapes, which acquire moral meaning before the perceptual matching task. Moreover, associations between shapes and different labels of moral characters are counter-balanced between participants, thus eliminating confounding effects by stimuli. Also, because we only used a few stimuli and they were repeatedly presented during the task, the results can not be explained by semantic priming (Unkelbach, Alves, & Koch, 2020), which is the center of the debate on previous results

(Firestone & Scholl, 2015, 2016a; Gantman & Bavel, 2015, 2016; Jussim et al., 2016). We examined whether participants' performance in the perceptual matching task was altered by the immediately acquired moral character of the shapes — in particular, whether the shapes associated with good or bad character are prioritized. We found a robust effect that shapes associated with good character are prioritized in the perceptual matching task. In a series of control experiments, we confirmed that moral content drove the prioritization effect, instead of other factors such as familiarity. In the subsequent experiments, we further tested whether the prioritization of moral character was caused by the valence of moral character alone or the interaction between valence and self-referential processing and found that only shapes associated with both good character and the self are prioritized, suggesting spontaneous moral self-referential as a novel mechanism underlying prioritization of good character in perceptual decision-making.

Disclosures

We reported all the measurements, analyses, and results in all the experiments in the current study. Participants whose overall accuracy was lower than 60% were excluded from analyses. Also, accurate responses with less than 200ms reaction times were excluded from the analysis. These excluded data can be found in the shared raw data files.

All the experiments reported were not pre-registered. Most experiments (1a ~ 4b, except experiment 3b) reported in the current study were first finished between 2013 to 2016 at Tsinghua University, Beijing, China. Participants in these experiments were recruited from the local community. To increase the sample size of experiments to 50 or more (Simmons, Nelson, & Simonsohn, 2013), we recruited additional participants from Wenzhou University, Wenzhou, China, in 2017 for experiments 1a, 1b, 4a, and 4b. Experiment 3b was finished at Wenzhou University in 2017 (See Table 1 for an overview of these experiments).

All participants received informed consent and were compensated for their time. These experiments were approved by the ethics board in the Department of Psychology, Tsinghua University.

General methods

Design and Procedure

This series of experiments used the social associative learning paradigm (or self-tagging paradigm, see Sui, He, and Humphreys (2012)), in which participants first learned the associations between geometric shapes and labels of different moral characters (e.g., in the first three studies, the triangle, square, and circle and Chinese words for "good person", "neutral person", and "bad person", respectively). The associations of shapes and labels were counterbalanced across participants. The paradigm consists of a brief learning stage and a test stage. During the learning stage, participants were instructed about the association between shapes and labels. Participants started the test stage with a practice phase to familiarize themselves with the task, in which they viewed one of the shapes above the fixation while one of the labels below the fixation and judged whether the shape and the label matched the association they learned. If the overall accuracy reached 60% or higher at the end of the practicing session, participants proceeded to the experimental task of the test stage. Otherwise, they finished another practices sessions until the overall accuracy was equal to or greater than 60%. The experimental task shared the same trial structure as in the practice.

Experiments 1a, 1b, 1c, 2, 5, and 6a were designed to explore and confirm the effect of moral character on perceptual matching. All these experiments shared a 2 (matching: match vs. nonmatch) by 3 (moral character: good vs. neutral vs. bad person) within-subject design. Experiment 1a was the first one of the whole series of studies, which aimed to examine the prioritization of moral character and found that shapes associated with good character were prioritized. Experiments 1b, 1c, and 2 were to confirm that it is the moral character that caused the effect. More specifically, experiment 1b used different Chinese words as labels to test whether the effect was contaminated by familiarity. Experiment 1c manipulated the moral character indirectly: participants first learned to associate different moral behaviors with different Chinese names, after remembering the association, they then associate the names with different shapes and finished the perceptual matching task. Experiment 2 further tested whether the way we presented the stimuli influence the prioritization of moral character, by sequentially presenting labels and shapes instead of simultaneous presentation. Note that a few participants in experiment 2 also participated in experiment 1a because we originally planned a cross-task comparison. Experiment 5 was designed to compare the prioritization of good character with other important social values (aesthetics and emotion). All social values had three levels, positive, neutral, and negative, and were associated with different shapes. Participants finished the associative learning task for different social values in different blocks, and the order of the social values was counterbalanced. Only the data from moral character blocks, which shared the design of experiment 1a, were reported here. Experiment 6a, which shared the same design as experiment 2, was an EEG experiment aimed at exploring the neural mechanism of the prioritization of good character. Only behavioral results of experiment 6a were reported here.

Experiments 3a, 3b, and 6b were designed to test whether the prioritization of good character can be explained by the valence effect alone or by an interaction between the valence effect and self-referential processing. To do so, we included self-reference as another within-subject variable. For example, experiment 3a extended experiment 1a into a 2 (matching: match vs. nonmatch) by 2 (reference: self vs. other) by 3 (moral character: good vs. neutral vs. bad) within-subject design. Thus, in experiment 3a, there were six conditions (good-self, neutral-self, bad-self, good-other, neutral-other, and bad-other) and six shapes (triangle, square, circle, diamond, pentagon, and trapezoids). Experiment 6b was an EEG experiment based on experiment 3a but presented the label and shape sequentially. Because of the relatively high working memory load (six label-shape pairs), participants finished experiment 6b in two days. On the first day, participants completed the perceptual matching task as a practice, and on the second day, they finished the task again while the EEG signals were recorded. We only focus on the first day's data here. Experiment 3b was designed to test whether the effect found in experiments 3a and 6b is robust if we separately present the self-referential trials and other-referential trials. That is, participants finished two different types of blocks: in the self-referential blocks, they only made matching judgments to shape-label pairs that related to the self (i.e., shapes and labels of good-self, neutral-self, and bad-self), in the other-referential blocks, they only responded to shape-label pairs that related to the other (i.e., shapes and labels of good-other, neutral-other, and bad-other).

Experiments 4a and 4b were designed to further test the interaction between valence and self-referential process in prioritization of good character. In experiment 4a, participants were instructed to learn the association between two shapes (circle and square) with two labels (self vs. other) in the learning stage. In the test stage, they were instructed only respond to the shape and label during the test stage. To test the effect of moral character, we presented the labels of moral character in the shapes and instructed participants to ignore the words in shapes when making matching judgments. In the experiment 4b, we reversed the role of self and moral character in the task: Participants learned associations between three labels (good-person, neutral-person, and bad-person) and three shapes (circle, square, and triangle) and made matching judgments about the shape and label of moral character, while words related to identity, "self" or "other", were presented within the shapes. As in 4a, participants were told to ignore the words inside the shape during the perceptual matching task.

Stimuli and Materials

We used E-prime 2.0 for presenting stimuli and collecting behavioral responses. Data were collected from two universities located in two different cities in China. Participants recruited from Tsinghua University, Beijing, finished the experiment individually in a dim-lighted chamber. Stimuli were presented on 22-inch CRT monitors and participants rested their chins on a brace to fix the distance between their eyes and the screen around 60 cm. The visual angle of geometric shapes was about $3.7° \times 3.7°$, the fixation cross is of $0.8° \times 0.8°$ visual angle at the center of the screen. The words were of $3.6° \times 1.6°$ visual angle. The distance between the center of shapes or images of labels and the fixation cross was of 3.5° visual angle. Participants from Wenzhou University, Wenzhou, finished the experiment in a group consisting of 3 ~ 12 participants in a dim-lighted testing room. They were instructed to finish the whole experiment independently. Also, they were told to start the experiment at the same time so that the distraction between participants was minimized. The stimuli were presented on 19-inch CRT monitors with the same set of parameters in E-prime 2.0 as in Tsinghua University, however, the visual angles could not be controlled because participants' chins were not fixed.

In most of these experiments, participants were also asked to fill out questionnaires after finishing the behavioral tasks. All the questionnaire data were open (see, dataset 4 in Liu et al., 2020). See Table 1 for a summary of information about all the experiments.

Data analysis

We used the `tidyverse` of r (see script `Load_save_data.r`) to preprocess the data. The data from all experiments were then analyzed using Bayesian hierarchical models.

We used the Bayesian hierarchical model (BHM, or Bayesian generalized linear mixed models, Bayesian multilevel models) to model the reaction time and accuracy data because BHM provided three advantages over the classic NHST approach (repeated measure

ANOVA or t-tests). First, BHM estimates the posterior distributions of parameters for statistical inference, therefore providing uncertainty in estimation (Rouder & Lu, 2005). Second, BHM, where generalized linear mixed models could be easily implemented, can use distributions that fit the distribution of real data instead of using the normal distribution for all data. Using appropriate distributions for the data will avoid misleading results and provide a better fitting of the data. For example, Reaction times are not normally distributed but are right skewed, and the linear assumption in ANOVAs is not satisfied (Rousselet & Wilcox, 2020). Third, BHM provides a unified framework to analyze data from different levels and different sources, avoiding information loss when we need to combine data from different experiments.

We used the `r` package `BRMs` (Bürkner, 2017), which used Stan (Carpenter et al., 2017) as the back-end, for the BHM analyses. We estimated the overall effect across experiments that shared the same experimental design using one model, instead of a two-step approach that was adopted in mini-meta-analysis (e.g., Goh, Hall, & Rosenthal, 2016). More specifically, a three-level model was used to estimate the overall effect of prioritization of good character, which included data from five experiments: 1a, 1b, 1c, 2, 5, and 6a. Similarly, a three-level HBM model is used for experiments 3a, 3b, and 6b. Results of individual experiments can be found in the supplementary results. For experiments 4a and 4b, which tested the implicit interaction between the self and good character, we used HBM for each experiment separately.

For questionnaire data, we only reported the subjective distance between different persons or moral characters in the supplementary results and did not analyze other questionnaire data, which are described in (Liu et al., 2020).

Response data. We followed previous studies (Hu, Lan, Macrae, & Sui, 2020; Sui et al., 2012) and used the signal detection theory approach to analyze the response data. More specifically, the match trials are treated as signals and non-match trials are noise. The sensitivity and criterion of signal detection theory are modeled through BHM (Rouder

& Lu, 2005).

We used the Bernoulli distribution for the signal detection theory. The probability that the jth subject responded "match" ($y_{ij} = 1$) at the ith trial p_{ij} is distributed as a Bernoulli distribution with parameter p_{ij}:

$$y_{ij} \sim Bernoulli(p_{ij})$$

The reparameterized value of p_{ij} is a linear regression of the independent variables:

$$\Phi(p_{ij}) = 0 + \beta_{0j} Valence_{ij} + \beta_{1j} IsMatch_{ij} * Valence_{ij}$$

where the probits (z-scores; Φ, "Phi") of ps is used for the regression.

The subjective-specific intercepts ($\beta_0 = -zFAR$) and slopes ($\beta_1 = d'$) are described by multivariate normal with means and a covariance matrix for the parameters.

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim N(\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}, \sum)$$

We used the following formula for experiments 1a, 1b, 1c, 2, 5, and 6a, which have a 2 (matching: match vs. non-match) by 3 (moral character: good vs. neutral vs. bad) within-subject design:

```
saymatch ~ 0 + Valence + Valence:ismatch + (0 + Valence +
Valence:ismatch | Subject) + (0 + Valence + Valence:ismatch |
ExpID_new:Subject) , family = bernoulli(link="probit")
```

in which the `saymatch` is the response data whether participants pressed the key corresponding to "match", `ismatch` is the independent variable of matching, `Valence` is the independent variable of moral character, `Subject` is the index of participants, and `Exp_ID_new` is the index of different experiments. Not that we distinguished data collected from two universities.

For experiments 3a, 3b, and 6b, an additional variable, i.e., reference (self vs. other), was included in the formula:

`saymatch ~ 0 + ID:Valence + ID:Valence:ismatch + (0 + ID:Valence + ID:Valence:ismatch | Subject) + (0 + ID:Valence + ID:Valence:ismatch | ExpID_new:Subject), family = bernoulli(link="probit")` in which the `ID` is the independent variable "reference", which means whether the stimulus was self-referential or other-referential.

Reaction times. We used log-normal distribution (https://lindeloev.github.io/shiny-rt/#34_(shifted)_log-normal) to model the RT data. This means that we need to estimate the posterior of two parameters: μ, and σ. μ is the mean of the `logNormal` distribution, and σ is the disperse of the distribution.

The reaction time of the jth subject on ith trial, y_{ij}, is log-normal distributed:

$$log(y_{ij}) \sim N(\mu_j, \sigma_j)$$

The parameter μ_j is a linear regression of the independent variables:

$$\mu_j = \beta_{0j} + \beta_{1j} * IsMatch_{ij} * Valence_{ij}$$

and the parameter σ_j does not vary with independent variables:

$$\sigma_j \sim HalfNormal()$$

The subjective-specific intercepts (β_{0j}) and slopes (β_{1j}) are described by multivariate normal with means and a covariance matrix for the parameters.

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim N(\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}, \sum)$$

The formula used for experiments 1a, 1b, 1c, 2, 5, and 6a, which have a 2 (matching: match vs. non-match) by 3 (moral character: good vs. neutral vs. bad) within-subject design, is as follows:

`RT_sec ~ 1 + Valence*ismatch + (Valence*ismatch | Subject) + (Valence*ismatch | ExpID_new:Subject), family = lognormal()` in which `RT_sec` is the reaction times data with the second as a unit. The other variables in this formula have the same meaning as the response data.

For experiments 3a, 3b, and 6b, which have a 2 by 2 by 3 within-subject design, the formula is as follows: `RT_sec ~ 1 + ID*Valence + (ID*Valence | Subject) + (ID*Valence | ExpID_new:Subject), family = lognormal()`

Note that for experiments 3a, 3b, and 6b, the three-level model for reaction times only included the matched trials to avoid divergence when estimating the posterior of the parameters.

Testing hypotheses.

Prioritization of moral character. We tested whether moral characters are prioritized by examining the population-level effects (also called fixed effect) of the three-level Bayesian hierarchical model of experiments 1a, 1b, 1c, 2, 5, and 6a. More specifically, we calculated the difference between the posterior distribution of the good/bad character and the neutral character and tested whether the 95% highest density intervals (HDIs) of the difference include zero. If the 95% highest density intervals do not include zero, we infer that there is a population-level difference between the conditions in the test, otherwise, we will infer that there is no evidence for such a difference. Note that for reaction times, we focused on the matched trials as in previous studies.

Modulation of self-referential processing. We tested the modulation effect of self-referential processing by examining the interaction between moral character and self-referential process for the three-level Bayesian hierarchical model of experiments 3a, 3b, and 6b. More specifically, we tested two possible explanations for the prioritization of good character: the valence effect alone or an interaction between the valence effect and the self-referential process. If the former is correct, then there will be no interaction between moral character and self-referential processing, i.e., the prioritization effect exhibits a similar pattern for both self- and other-referential conditions. On the other hand, if the spontaneous self-referential processing account is true, then there will be an interaction between the two factors, i.e., the prioritization effect exhibits different patterns for self- and other-referential conditions.

Spontaneous binding between the self and good character. For data from experiments 4a and 4b, we further examined whether the self-referential processing for moral characters is spontaneous (i.e., whether the good character is spontaneously bound with the self). For experiment 4a, if there exists a spontaneous binding between self and good character, there should be an interaction between moral character and self-referential processing, e.g., the task-irrelevant moral words either facilitate or slows down the response to self- or other-referential conditions. For experiment 4b, if there exists a spontaneous binding between self and good character, then, there will be a self-other difference for some moral character conditions but not for other moral character conditions.

Results

Prioritization of good character

To test whether moral characters are prioritized, we modeled data from experiments 1a, 1b, 1c, 2, 5, and 6a with three-level Bayesian hierarchical models. All these experiments shared similar designs and can be used for testing the prioritization effect of moral character. The valid and unique sample size is 192. Note that for both experiments 1a and 1b, two datasets were collected at different time points and locations, thus we treated them as independent samples. Here we only reported the population-level results of three-level Bayesian models, the detailed results of each experiment can be found in supplementary materials.

For the *d* prime, results from the Bayesian model revealed a robust effect of moral character. Shapes associated with good characters ("good person", "kind person" or a name associated with good behaviors) have higher sensitivity (median = 2.51, 95% HDI = [2.23 2.78]) than shapes associated with neutral characters (median = 2.19, 95% HDI = [1.88 2.50]), $median_{diff}$ = 0.31, 95% HDI [0.00 0.64], but we did not find differences between shapes associated with bad characters (median = 2.25, 95% HDI = [1.94 2.55]) and neutral character, $median_{diff}$ = 0.05, 95% HDI [-0.28 0.39].

The results from reaction times data also found a robust effect of moral character for both match trials (see figure 1 C) and nonmatch trials (**see supplementary materials**). For match trials, shapes associated with good characters were faster (median = 579.03 ms, 95% HDI = [500.20 660.89]) than shapes associated with neutral characters (median = 623.59 ms, 95% HDI = [542.83 710.82]), $median_{diff}$ = -44.19, 95% HDI [-59.85 -30.36]. We also found that RTs to shapes associated with bad characters (median = 640.86 ms, 95% HDI = [561.22 729.99]) were slower as compared to the neutral character, $median_{diff}$ = 16.85, 95% HDI [2.82 30.10].

For the nonmatch trials, we found the advantage of good character: Shapes associated with good characters (median = 654.16 ms, 95% HDI = [573.12 742.91]) were faster than shapes associated with neutral characters (median = 671.81 ms, 95% HDI = [588.33 762.65]), $median_{diff}$ = -17.72 ms, 95% HDI [-24.58 -11.19]. Similarly, the shapes associated with bad characters (median = 676.93 ms, 95% HDI = [590.23 765.67]) were slower than shapes associated with neutral characters, $median_{diff}$ = 5.20 ms, 95% HDI [-0.04 10.66], but the effect size was smaller than the match trials.

Modulation effect self-referential processing

To test the modulation effect of self-referential processing, we also modeled data from three experiments (3a, 3b, and 6b) with three-level Bayesian models. These three experiments included 108 unique participants. We focused on the population-level effect of the interaction between self-referential processing and moral valence. Also, we examined the differences of differences, i.e., how the differences between good/bad characters and the neutral character under the self-referential conditions differ from that under other-referential conditions. The detailed results of each experiment can be found in supplementary materials.

For the *d* prime, we found an interaction between the moral valence and self-referential processing: the good-neutral differences are larger for the self-referential condition than for the other-referential condition ($median_{diff}$ = 0.52; 95% HDI = [-1.54 2.72]). However, this is not the case for the bad-neutral differences ($median_{diff}$ = -0.01; 95% HDI = [-1.55 2.37]). Further analyses revealed that the prioritization effect of good character (as compared to neutral) only appeared for self-referential conditions but not other-referential conditions. The estimated *d* prime for good-self was greater than neutral-self ($median_{diff}$ = 0.57; 95% HDI = [-0.93 2.88]), *d* prime for good-self was also greater than good-other condition ($median_{diff}$ = 0.23; 95% HDI = [-1.31 1.87]). The differences between bad-self and neutral-self, good-other and neutral-other, and bad-other and neutral-other are all centered around zero (see Figure 2, B, D).

For the RTs of matched trials, we also found an interaction between moral valence and self-referential processing: the good-neutral differences were different for the self- and other-referential conditions ($median_{diff}$ = -155.11; 95% HDI = [-755.36 395.38]). However, this was not the case for bad-neutral differences ($median_{diff}$ = -55.63; 95% HDI = [-604.70 561.00]). Further analyses revealed a robust good-self prioritization effect as compared to neutral-self ($median_{diff}$ = -57.91; 95% HDI = [-224.43 41.14]) and good-other ($median_{diff}$ = -107.21; 95% HDI = [-369.05 92.95]) conditions. Similar to the results of *d'*, we found that participants responded slower for both good character and bad character than for the neutral character when they referred to others. See Figure 2.

These results suggested that the prioritization of good character is not solely driven by the valence of moral character. Instead, the self-referential processing modulated the prioritization of good character: good character was prioritized only when it was self-referential. When the moral character was other-referential, responses to both good and bad characters were slowed down.

Spontaneous binding between the good character and the self

Experiments 4a and 4b were designed to test whether the good character and self-referential processing bind together spontaneously. Because these two experiments have different experimental designs, we model their data separately.

In experiment 4a, where "self" vs. "other" were task-relevant and moral character were task-irrelevant, we found the "self" conditions performed better than the "other" conditions for both *d* prime and reaction times. This pattern is consistent with previous studies (e.g., Sui et al. (2012)).

More importantly, we found evidence, albeit weak, that task-irrelevant moral character also played a role. For shapes associated with "self", *d'* was greater when shapes had a good character inside (median = 2.83, 95% HDI [2.63 3.01]) than shapes that have neutral character (median = 2.74, 95% HDI [2.58 2.95]) or bad character (median = 2.76, 95% HDI [2.56 2.95]), but this is not the case for self-referential shapes with bad character and neutral character inside. For shapes associated with "other", the pattern reversed: *d* prime was smaller when shapes had a good character inside (median = 1.87, 95% HDI [1.71 2.04]) than had neutral (median = 1.96, 95% HDI [1.80 2.14]) or bad character (median = 1.98, 95% HDI [1.79 2.17]) inside. See Figure 3.

A similar pattern was found for RTs in matched trials. For the "self" condition, when a good character was presented inside the shapes, the RTs (median = 641, 95% HDI [623 662]) were faster than when a neutral character (median = 649, 95% HDI [631 668]) or bad character (median = 648, 95% HDI [629 667]) were inside. This effect was reversed for the "other" condition: RTs for shapes associated with good character inside (median = 733, 95% HDI [711 755]) were slower than those with neutral character (median = 722, 95% HDI [702 741]) or bad character (median = 719, 95% HDI [696 740]) inside.

In experiment 4b, where moral characters were task-relevant and "self" vs "other" were task-irrelevant, we found a main effect of moral character: performance for shapes associated with good characters was better than other-related conditions on both *d'* and reaction times. This pattern, again, shows a robust prioritization effect of good character.

Most importantly, we found evidence that task-irrelevant labels, "self" or "other", also played a role. For shapes associated with good character, the *d* prime was greater when shapes had a "self" inside than with "other" inside ($mean_{diff}$ = 0.14, 95% credible intervals [-0.02, 0.31], BF = 12.07), but this effect did not occur when the target shape where associated with "neutral" ($mean_{diff}$ = 0.04, 95% HDI [-.11, .18]) or "bad" person ($mean_{diff}$ = -.05, 95% HDI[-.18, .09]).

The same trend appeared for the RT data. For shapes associated with good character, having a "self" inside shapes reduced the reaction times as compared to having an "other" inside the shapes ($mean_{diff}$ = -55 ms, 95% HDI [-75, -35]), but this effect did not occur when the shapes were associated neutral ($mean_{diff}$ = 10, 95% HDI [1, 20]) or bad ($mean_{diff}$ = 5, 95% HDI [-16, 27]) person, See Figure 3.

Discussion

Across nine experiments, we explored the prioritization effect of moral character and the underlying mechanism by a combination of social associative learning and perceptual matching task. First, we found a robust effect that good character was prioritized in the shape-label matching task across five experiments. Second, across three experiments, we found that the prioritization of good character was not solely driven by moral valence itself, i.e., "good" vs "bad". Instead, this effect was modulated by self-referential processing: prioritization only occurred when moral characters are self-referential. Finally, the prioritization of the combination of good character and self occurred, albeit weak, even when either the self- or character-related information was irrelevant to the experimental task (experiment 4a and 4b). In contrast, performance to the combination of good character and "other", explicitly or implicitly, was worse than the combination of neutral character and "other". Together, these results highlighted the importance of the self in perceiving information related to moral characters, suggesting a spontaneous self-referential process when making perceptual decision-making for moral characters. These results are in line with a growing literature on the social and relational nature of perception (Xiao, Coppin, and Bavel (2016); Freeman, Stolier, and Brooks (2020); hafri_perception_2021) and deepened our understanding of mechanisms of perceptual decision-making of moral information.

The current study provided robust evidence for the prioritization of good character in perceptual decision-making. The existence of the effect of moral valence on perception has been disputed. For instance, (E. Anderson et al., 2011) reported that faces associated with bad social behavior capture attention more rapidly, however, an independent team failed to replicate the effect (Stein et al., 2017). Another study by Gantman and Van Bavel (2014) found that moral words are more likely to be judged as words when it was presented subliminally, however, this effect may be caused by semantic priming instead of morality (Firestone & Scholl, 2015; Jussim et al., 2016). In the current study, we found the prioritization effect across five experiments, the sample size of individual experiments and combined provide strong evidence for the existence of the effect. Moreover, the associative learning task allowed us to eliminate the semantic priming effect for two reasons. First, associations between shapes and moral characters were acquired right before the perceptual matching task, semantic priming from pre-existed knowledge was impossible. Second, there were only a few pairs of stimuli were used and each stimulus represented different conditions, making it impossible for priming between trials. Importantly, a series of control experiments (1b, 1c, and 2) further excluded other confounding factors such as familiarity, presenting sequence, or words-based associations, suggesting that it was the moral content that drove the prioritization of good character.

The robust prioritization of good character found in the current study was incongruent with previous moral perception studies, which usually reported a negativity effect, i.e., information related to bad character is processed preferentially (E. Anderson et al., 2011; Eiserbeck & Abdel Rahman, 2020). This discrepancy may be caused by the experimental task: while in many previous moral perception studies, the participants were asked to detect the existence of a stimulus, the current task asked participants to recognize a pattern. In other words, previous studies targeted early stages of perception while the current task focused more on decision-making at a relatively later stage of information processing. This discrepancy is consistent with the pattern found in studies with emotional stimuli (Pool, Brosch, Delplanque, & Sander, 2016).

We expanded previous moral perception studies by focusing on the agent who made the perceptual decision-making and examined the interaction between moral valence and self-referential processing. Our results revealed that prioritization of good character is modulated by self-referential processing: the good character was prioritized when it was related to the "self", even when the self-relatedness was task-irrelevant. By contrast, good character information was not prioritized when it was associated with "other". The modulation effect of self-referential processing was large when the relationship between moral character and the self was explicit, which is consistent with previous studies that only positive aspects of the self are prioritized (Hu et al., 2020). More importantly, the effect persisted when the relationship between moral character and self-information was implicit, suggesting spontaneous self-referential processing when both pieces of information were presented. A possible explanation for this spontaneous self-referential of good character is that the positive moral self-view is central to our identity (Freitas, Cikara, Grossmann, & Schlegel, 2017; Strohminger, Knobe, & Newman, 2017) and the motivation to maintain a moral self-view influences how we perceive (e.g., Ma & Han, 2010) and remember (e.g., Carlson, Maréchal, Oud, Fehr, & Crockett, 2020; Stanley, Henne, & De Brigard, 2019).

Although the results here revealed the prioritization of good character in perceptual decision-making, we did not claim that the motivation of a moral self-view *penetrates* perception. The perceptual decision-making process involves processes more than just encoding the sensory inputs. To fully account for the nuance of behavioral data and/or related data collected from other modules (e.g., Sui, He, Golubickis, Svensson, & Neil Macrae, 2023), we need computational models and an integrative experimental approach (Almaatouq et al., 2022). For example, sequential sampling models suggest that, when making a perceptual decision, the agent continuously accumulates evidence until the amount of evidence passed a threshold, then a decision is made (Chuan-Peng et al., 2022; Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff, Smith, Brown, & McKoon, 2016). In these models, the evidence, or decision variable, can accumulate from both sensory information but also memory (Shadlen & Shohamy, 2016). Recently, applications of sequential sample models to perceptual matching tasks also suggest that different processes may contribute to the prioritization effect of self (Golubickis et al., 2017) or good self (Hu et al., 2020). Similarly, reinforcement learning models also revealed that the key difference between self- and other-referential learning lies in the learning rate (Lockwood et al., 2018). These studies suggest that computational models are needed to disentangle the cognitive processes underlying the prioritization of good character.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). *Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences.* 1–55. https://doi.org/10.1017/S0140525X22002874

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, *108*(25), 10367–10371. https://doi.org/10.1073/pnas.1104047108

Anderson, E., Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The visual impact of gossip. *Science*, *332*(6036), 1446–1448. https://doi.org/10.1126/science.1201574

Atlas, L. Y., Dildine, T. C., Palacios-Barrios, E. E., Yu, Q., Reynolds, R. C., Banker, L. A., … Pine, D. S. (2022). Instructions and experiential learning have similar impacts on pain and pain-related brain responses but produce dissociations in value-based reversal learning. *eLife*, *11*, e73353. https://doi.org/10.7554/eLife.73353

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan [Journal Article]. *Journal of Statistical Software; Vol 1, Issue 1 (2017)*. Retrieved from https://www.jstatsoft.org/v080/i01 http://dx.doi.org/10.18637/jss.v080.i01

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 2100. https://doi.org/10.1038/s41467-020-15602-4

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language [Journal Article]. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Chuan-Peng, H., Geng, H., Zhang, L., Fengler, A., Frank, M., & Zhang, R.-Y. (2022). *A Hitchhiker's Guide to Bayesian Hierarchical Drift-Diffusion Modeling with dockerHDDM*. PsyArXiv. https://doi.org/10.31234/osf.io/6uzga

Deltomme, B., Mertens, G., Tibboel, H., & Braem, S. (2018). Instructed fear stimuli bias visual attention. *Acta Psychologica*, *184*, 31–38. https://doi.org/10.1016/j.actpsy.2017.08.010

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, *8*(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

Eiserbeck, A., & Abdel Rahman, R. (2020). Visual consciousness of faces in the attentional blink: Knowledge-based effects of trustworthiness dominate over appearance-based impressions. *Consciousness and Cognition*, *83*, 102977. https://doi.org/10.1016/j.concog.2020.102977

Ellemers, N. (2018). Morality and social identity. In M. van Zomeren & J. F. Dovidio (Eds.), *The oxford handbook of the human essence* (pp. 147–158). New York, NY, US: Oxford University Press.

Firestone, C., & Scholl, B. J. (2015). Enhanced visual awareness for morality and pajamas? Perception vs. Memory in ”top-down” effects. *Cognition*, *136*, 409–416. https://doi.org/10.1016/j.cognition.2014.10.014

Firestone, C., & Scholl, B. J. (2016a). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, *39*, e229. https://doi.org/10.1017/S0140525X15000965

Firestone, C., & Scholl, B. J. (2016b). “Moral Perception” Reflects Neither Morality Nor Perception. *Trends in Cognitive Sciences*, *20*(2), 75–76. https://doi.org/10.1016/j.tics.2015.10.006

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*(1). https://doi.org/10.1146/annurev-psych-122414-033645

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Chapter five - dynamic interactive theory as a domain-general account of social perception. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 61, pp. 237–287). Academic Press. https://doi.org/10.1016/bs.aesp.2019.09.005

Freitas, J. D., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences*, *21*(9), 634–636. https://doi.org/10.1016/j.tics.2017.05.009

Gantman, A. P., & Bavel, J. J. V. (2015). Moral Perception. *Trends in Cognitive Sciences*, *19*(11), 631–633. https://doi.org/10.1016/j.tics.2015.08.004

Gantman, A. P., & Bavel, J. J. V. (2016). See for Yourself: Perception Is Attuned to Morality. *Trends in Cognitive Sciences*, *20*(2), 76–77. https://doi.org/10.1016/j.tics.2015.12.001

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*(1), 22–29. https://doi.org/10.1016/j.cognition.2014.02.007

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Golubickis, M., Falben, J. K., Sahraie, A., Visokomogilski, A., Cunningham, W. A., Sui, J., & Macrae, C. N. (2017). Self-prioritization and perceptual matching: The effects of temporal construal. *Memory & Cognition*, *45*(7), 1223–1239. https://doi.org/10.3758/s13421-017-0722-3

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Hu, C.-P., Lan, Y., Macrae, C. N., & Sui, J. (2020). Good me bad me: Does valence influence self-prioritization during perceptual decision-making? *Collabra: Psychology*, *6*(1), 20. https://doi.org/10.1525/collabra.301

Juechems, K., & Summerfield, C. (2019). Where does value come from? *Trends in Cognitive Sciences*, *23*(10), 836–850. https://doi.org/10.1016/j.tics.2019.07.012

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology*, *66*, 116–133. https://doi.org/10.1016/j.jesp.2015.10.003

Liu, Q., Wang, F., Yan, W., Peng, K., Sui, J., & Hu, C.-P. (2020). Questionnaire data from the revision of a chinese version of free will and determinism plus scale. *Journal of Open Psychology Data*, *8*(1), 1. https://doi.org/10.5334/jopd.49/

Lockwood, P. L., Wittmann, M. K., Apps, M. A. J., Klein-FlÃŒgge, M. C., Crockett, M. J., Humphreys, G. W., & Rushworth, M. F. S. (2018). Neural mechanisms for learning self and other ownership. https://doi.org/10.1038/s41467-018-07231-9

Ma, Y., & Han, S. (2010). Why we respond faster to the self than to others? An implicit positive association theory of self-advantage during implicit face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 619–633. https://doi.org/10.1037/a0015797

Ohman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, *80*(3), 381–396. https://doi.org/10.1037/0022-3514.80.3.381

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. https://doi.org/10.1037/bul0000026

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Rouder, J. N., & Lu, J. (2005). An introduction to bayesian hierarchical models with an application in the theory of signal detection [Journal Article]. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

https://doi.org/10.3758/bf03196750

Rousselet, G. A., & Wilcox, R. R. (2020). Reaction times and other skewed distributions: Problems with the mean and the median. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2019.1630

Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*(5), 927–939. https://doi.org/10.1016/j.neuron.2016.04.036

Shore, D. M., & Heerey, E. A. (2013). Do social utility judgments influence attentional processing? *Cognition*, *129*(1), 114–122. https://doi.org/10.1016/j.cognition.2013.06.011

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after p-hacking* [Conference Proceedings]. https://doi.org/10.2139/ssrn.2205186

Stanley, M. L., Henne, P., & De Brigard, F. (2019). Remembering moral and immoral actions in constructing the self. *Memory & Cognition*, *47*(3), 441–454. https://doi.org/10.3758/s13421-018-0880-y

Stein, T., Grubb, C., Bertrand, M., Suh, S. M., & Verosky, S. C. (2017). No impact of affective person knowledge on visual awareness: Evidence from binocular rivalry and continuous flash suppression. *Emotion*, *17*(8), 1199–1207. https://doi.org/10.1037/emo0000305

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self: *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691616689495

Sui, J., He, X., Golubickis, M., Svensson, S. L., & Neil Macrae, C. (2023). Electrophysiological correlates of self-prioritization. *Consciousness and Cognition*, *108*, 103475. https://doi.org/10.1016/j.concog.2023.103475

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of*

Experimental Psychology: Human Perception and Performance, *38*(5), 1105–1117. https://doi.org/10.1037/a0029792

Sui, J., & Humphreys, G. W. (2015). The Integrative Self: How Self-Reference Integrates Perception and Memory. *Trends in Cognitive Sciences*, *19*(12), 719–728. https://doi.org/10.1016/j.tics.2015.08.015

Sui, J., & Rotshtein, P. (2019). Self-prioritization and the attentional systems. *Current Opinion in Psychology*, *29*, 148–152. https://doi.org/10.1016/j.copsyc.2019.02.010

Unkelbach, C., Alves, H., & Koch, A. (2020). Chapter three - negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 62, pp. 115–187). Academic Press. https://doi.org/10.1016/bs.aesp.2020.04.005

Xiao, Y. J., Coppin, G., & Bavel, J. J. V. (2016). Perceiving the world through group-colored glasses: A perceptual model of intergroup relations. *Psychological Inquiry*, *27*(4), 255–274. https://doi.org/10.1080/1047840X.2016.1199221

Table 1

Information about all experiments.

ExpID	Time	Location	N	n.of.trials	Self.ref	Stim.for.Morality	Presenting.order
Exp_1a_1	2014-04	Beijing	38 (35)	60	NA	words	Simultaneously
Exp_1a_2	2017-04	Wenzhou	18 (16)	60	NA	words	Simultaneously
Exp_1b_1	2014-10	Beijing	39 (27)	NA	NA	words	Simultaneously
Exp_1b_2	2017-04	Wenzhou	33 (25)	NA	NA	words	Simultaneously
Exp_1c	2014-10	Beijing	23 (23)	NA	NA	descriptions	Simultaneously
Exp_2	2014-05	Beijing	35 (34)	NA	NA	words	Sequentially
Exp_3a	2014-11	Beijing	38 (35)	NA	explicit	words	Simultaneously
Exp_3b	2017-04	Wenzhou	61 (56)	NA	explicit	words	Simultaneously
Exp_4a_1	2015-06	Beijing	32 (29)	NA	implicit	words	Simultaneously
Exp_4a_2	2017-04	Wenzhou	32 (30)	NA	implicit	words	Simultaneously
Exp_4b_1	2015-10	Beijing	34 (32)	NA	implicit	words	Simultaneously
Exp_4b_2	2017-04	Wenzhou	19 (13)	NA	implicit	words	Simultaneously
Exp_5	2016-01	Beijing	43 (38)	NA	NA	words	Simultaneously
Exp_6a	2014-12	Beijing	24 (24)	NA	NA	words	Sequentially
Exp_6b	2016-01	Beijing	23 (22)	NA	explicit	words	Sequentially
Exp_7a	2016-07	Beijing	35 (29)	NA	explicit	words	Simultaneously
Exp_7b	2018-05	Beijing	46 (42)	NA	explicit	words	Simultaneously

Note. Stim of Morality = How moral character was manipulated; Presenting order = how shapes & labels were presented. The data from experiments 7a & 7b, which were reported in Hu et al (2020), are only included in the meta-analysis in supplementary materials.

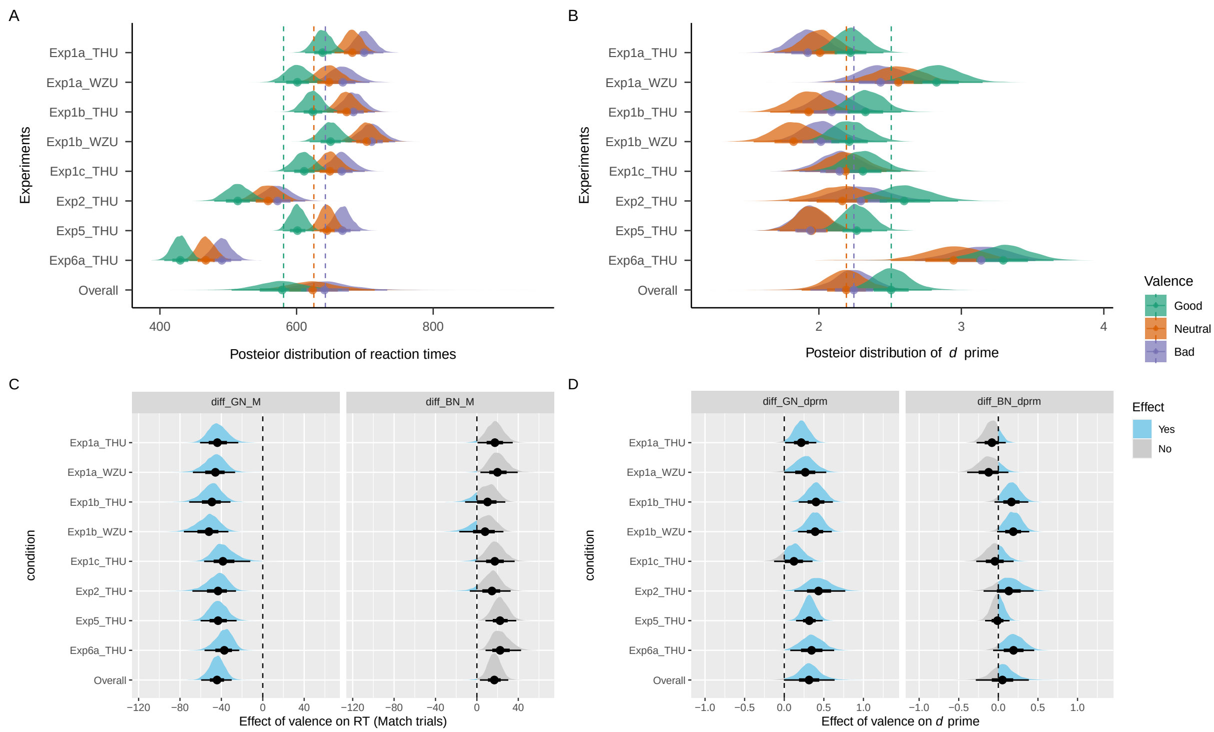

Figure 1. Effect of moral character on RT and d'

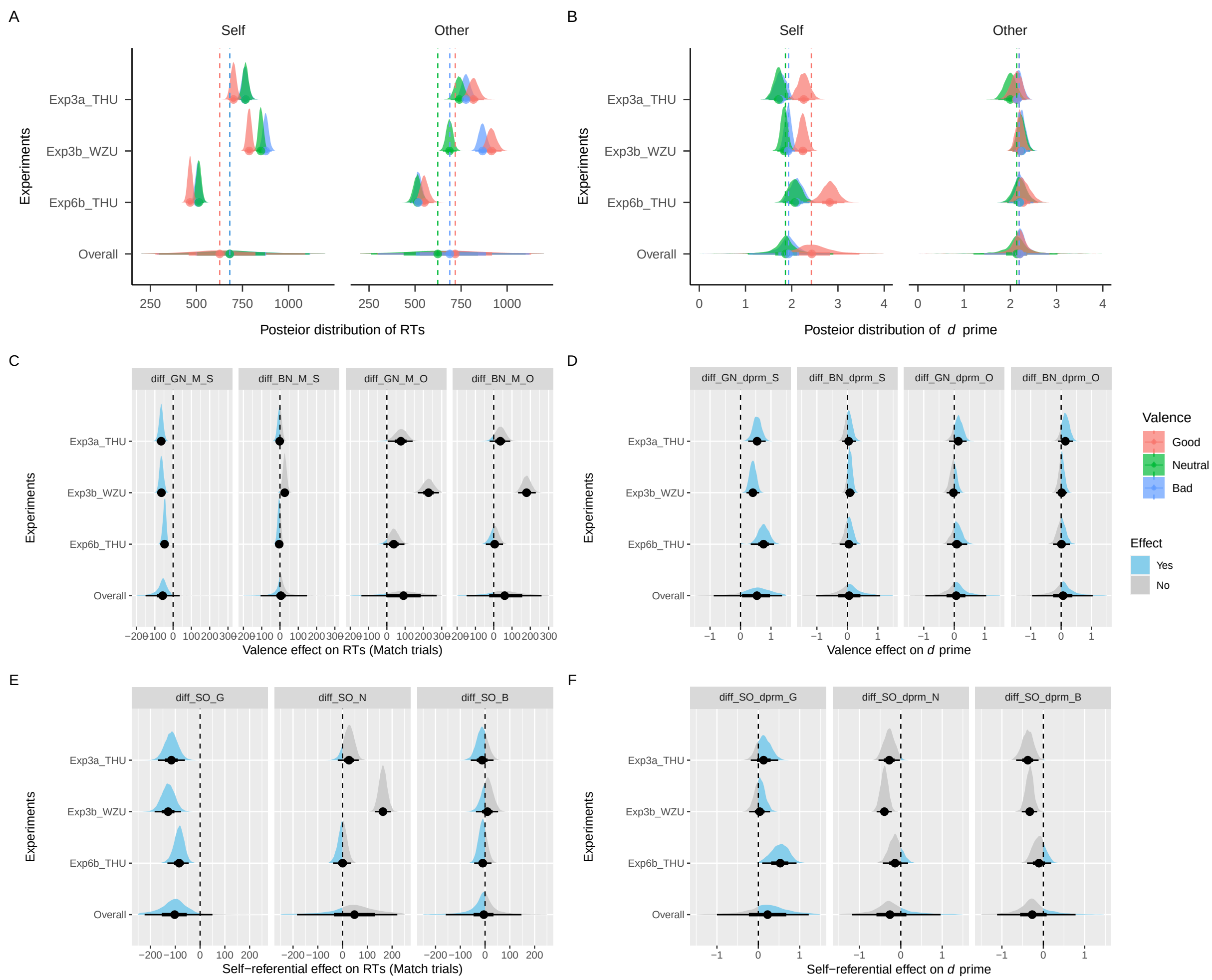

Figure 2. Interaction between moral character and self-referential

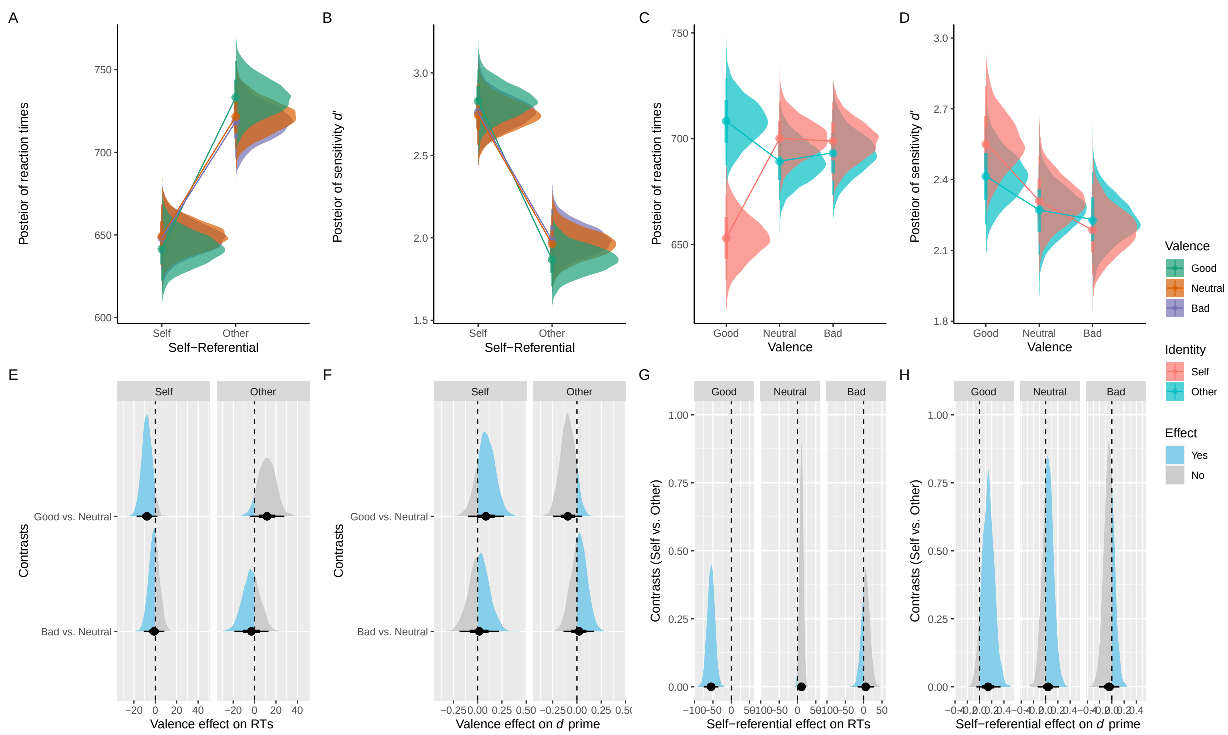

Figure 3. Experiment 4: Implicit binding between good character and the self.