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 geometric 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 or self-reference. Data from three experiments (N = 108) demonstrated that the prioritization effect of good character was robust only 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

Alternative title: Self-relevance modulates the prioritization of the good character in perceptual matching

Introduction

[quotes about moral character]

Is moral information prioritized in perception? This question 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 [XX], rewards [XX], or self-related stimuli [XX]. Indeed, previous studies reported bad characters are prioritized in visual processing (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 [XX]. 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 [XXX]. The positive bias toward moral information, on the other hand, is explained by value-based attention [XXX]. 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__tics__2015?**). 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 was not only used in studying the self-relevance effect but also other factors such as aversive stimuli [XX] and rewards [XX]. 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, 2016; Gantman & Bavel, 2015, 2016; Jussim, Crawford, Anglin, Stevens, & Duarte, 2016; **firestone_cognition_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 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 lower than 60% were excluded from analysis. Also, the 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 in Tsinghua University, Beijing, China. Participants in these experiments were recruited in the local community. To increase the sample size of experiments to 50 or more (Simmons, Nelson, & Simonsohn, 2013), we recruited additional participants in Wenzhou University, Wenzhou, China in 2017 for experiment 1a, 1b, 4a, and 4b. Experiment 3b was finished in Wenzhou University in 2017 (See Table S1 for overview of these experiments).

All participants received informed consent and compensated for their time. These experiments were approved by the ethic 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 validate 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 revealed a prioritization of good character. Experiments 1b, 1c, and 2 were then designed

to confirm that it is moral character that caused the effect. More specifically, experiment 1b used different Chinese words as labels to test whether the effect only occurred with specific words. 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 effect size of moral character and other important social evaluative dimensions (aesthetics and emotion). Different social evaluative dimensions were implemented in different blocks, only the data from moral character blocks, which shared the design of experiment 1a, were included in the three-level models. Experiment 6a, which shared the same design as experiment 2, was an EEG experiment aimed at exploring the neural correlates of the effect. Only the behavioral results of experiment 6a were included in the analysis 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 directly 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), experiment 6b was conducted in two days: on the first day participants finished 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 separate 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 mechanism of the self-referential process in the 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 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

E-prime 2.0 was used for presenting stimuli and collecting behavioral responses. The experiments were conducted in 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 rest 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 $3.7^\circ \times 3.7^\circ$, the fixation cross is of $0.8^\circ \times 0.8^\circ$ visual angle at the center of the screen. The words were of $3.6^\circ \times 1.6^\circ$ visual angle. The distance between the center of shapes or images of labels and the fixation cross was of 3.5° visual angle. Participants recruited from Wenzhou University, Wenzhou, finished the experiment in a group consisting of $3 \sim 12$ participants in a dim-lighted testing room. They were required to finish the whole experiment independently. Also, they were instructed to start the experiment simultaneously, 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 S1 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, 2019). 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 backend, 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 good-person on perceptual matching, which include data from five experiments: 1a, 1b, 1c, 2, 5, and 6a. Similarly, a three-level HBM model is used for all experiments that tested the modulation effect of self-referential processing on the prioritization of good character, including experiments 3a, 3b, and 6b. Results of the individual experiment 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.

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.

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

In `BRMs`, we used the following formula for the first set of experiments, which have a 2 (matching: match vs. non-match) by 3 (moral character: good vs. neutral vs. bad) within-subject design, i.e., experiment 1a, 1b, 1c, 2, 5, and 6a:

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

For experiments 3a, 3b, and 6b, an additional variable is 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")
```

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 is log-normal distributed:

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

y_{ij} is the RT of the ith trial of the jth participants.

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

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

$$\sigma_j \sim HalfNormal()$$

The formula used for modeling the data from experiments with a 2 by 3 within-subject design is as follows:

```
RT_sec ~ 1 + Valence*ismatch + (Valence*ismatch | Subject), family = lognormal()
```

For experiments with a 2 by 2 by 3 within-subject design, the formula is as follows:

```
RT_sec ~ 1 + ID*Valence*ismatch + (ID*Valence*ismatch | Subject), family = lognormal()
```

As for the *d*', we expanded the RT model to three-level models and the formulas used for three-level models are as follows:

```
RT_sec ~ 1 + Valence*ismatch + (Valence*ismatch | Subject) + (Valence*ismatch | ExpID_new:Subject), family = lognormal()
```

or

```
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.

Prioritization of moral character. We tested whether moral characters are prioritized by examining the population-level effects from the three-level Bayesian hierarchical model of d' and reaction times from 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.

Interaction between valence and self-referential processing. We tested the interaction between valence and self-referential processing by examining the interaction between moral character and self-referential process, which included results from experiments 3a, 3b, and 6b. Using three-level models, we tested two possible explanations for the prioritization of good character: the valence effect or the interaction between self-referential processing and valence. If the valence effect model is correct, then there will be no interaction between moral character and self-referential processing, i.e., the prioritization effect exhibit a similar pattern for both self- and other-referential conditions. On the other hand, if the interaction model is true, then there will be an interaction between the two factors, i.e., the prioritization effect exhibit different patterns for self- and other-referential conditions.

Spontaneous interaction between the self and good character. Based on data from experiments 4a and 4b, we further examined whether the self-referential processing for moral characters is spontaneous. The design of experiments 4a and 4b are complementary. 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 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 should also be an interaction between the two, e.g., there will be a self-other difference for some moral character conditions but not for other moral character conditions.

We only reported the subjective distance between different persons in the supplementary results and did not analyze other questionnaire data described in (Liu et al., 2020).

Results

Prioritization of good character related information

In this part, we report results from five experiments that tested whether an associative learning task, including 192 participants. Note that for both experiment 1a and 1b, there were two independent samples with different equipment, trials numbers and testing situations. Therefore, we modeled them as independent samples. These five experiments revealed a robust effect of moral character on perceptual matching task.

For the d prime, we found robust effect of moral character. Shapes associated with good character ("good person", "kind person" or a name associated with morally good behavioral history) has higher sensitivity (median = 2.51, 95% HDI = [2.23 2.78]) than shapes associated with neutral character (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 character (median = 2.25, 95% HDI = [1.94 2.55]) and neutral character, $median_{diff}$ = 0.05, 95% HDI [-0.28 0.39].

For the reaction times, we also found 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 character has faster responses (median = 579.03 ms, 95% HDI = [500.20 660.89]) than shapes associated with neutral character (median = 623.59 ms, 95% HDI = [542.83 710.82]), $median_{diff}$ = -44.19, 95% HDI [-59.85 -30.36].

We also found that the responses to shapes associated with bad character (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]. See Figure 1.

For the nonmatch trials, we also found the advantage of good character: Shapes associated with good character (median = 654.16 ms, 95% HDI = [573.12 742.91]) are faster than shapes associated with neutral (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 character (median = 676.93 ms, 95% HDI = [590.23 765.67]) was responded slower than shapes associated with neutral character, $median_{diff}$ = 16.85 ms, 95% HDI [2.82 30.10], but the effect size was smaller than the match trials (**see supplementary materials**).

Self-referential process modulates prioritization of good character

In this part, we report results from three experiments (3a, 3b, and 6b) that aimed at testing whether the moral valence effect found in the previous experiments is modulated by self-referential processes. These three experiments included data from 108 unique participants.

Given that we already found a prioritization effect of good character and a slow-down effect of bad character, here we focused on the whether such effect modulated by self-referential factor or purely driven by valence. To test the modulation effect, our results focused on the differences between good/bad character with neutral character for self-referential and other-referential separately, also, we compared the differences between the difference, i.e., how did differences between good and bad characters under the self-referential conditions differ from that under other-referential conditions. The details of individual studies can be found in supplementary materials.

For the d prime, we found that an interaction between moral character effect and

self-referential, the good-neutral differences is larger for 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 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 ($median_{diff}$ = ; 95% HDI = [ ]). The differences between bad-self and neutral-self, good-other and neutral-other, bad-other and neutral-other are all centered around zero (see Figure 2, B, D).

For the RTs part, we also found the interaction between moral character and self-referential, the good-neutral differences were different for the self- and other-referential conditions ($median_{diff}$ = -155.11; 95% HDI = [-755.36 395.38]). Again, 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. As the results of *d'*, we found that both good character and bad character were responded slower than neutral character for other-referential conditions. See Figure 2.

These results suggested that the prioritization of good character is not driven by valence of moral character. Instead, the self-referential processing modulated the prioritization of good character: the good character was prioritized only when it was self-referential. When the moral character was other-referential, good character has similar performance as bad character.

Spontaneous binding between the good character and the self

Two studies further tested whether the binding between self and good character happen even when these two piece of information are separated and only one of them is task-relevant. We are interested in testing whether the task-relevance modulated the effect observed in previous experiment.

In experiment 4a, where self- and other-referential were task-relevant and moral character are task-irrelevant. We found self-related conditions were performed better than other-related conditions, on 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 an role. For shapes associated with self, d' was greater when shapes had a good character inside the shape (median = 2.83, 95% HDI [2.63 3.01]) than shapes that have neutral character (median = 2.74, 95% HDI [2.58 2.95], BF = 4.4) or bad character (median = 2.76, 95% HDI [2.56 2.95], 3.1), but we did not found difference between shapes with bad character and neutral character inside for the self-referential shapes. For shapes associated with other, the results of d' revealed a reversed pattern to the self-referential condition: 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.

The same pattern was found for RTs. For self-referential condition, when good character was presented as a task-irrelevant stimuli, the responds (median = 641, 95% HDI [623 662]) were faster than when neutral character (median = 649, 95% HDI [631 668]) or bad character (median = 648, 95% HDI [628 667]) were inside. This effect was reversed for other-referential condition: shapes associated with other with good character inside (median = 733, 95% HDI [711 754]) were slower than with neutral character (median = 721, 95% HDI [702 741]) or bad character (median = 718, 95% HDI [696 740]) inside.

In experiment 4b, moral character was the task-relevant factor, and we found that there were main effect of moral character: shapes associated with good character were performed better than other-related conditions, on both *d'* and reaction times.

Most importantly, we found evidence that task-irrelevant self-referential process also played an role. For shapes associated with good person, the *d* prime was greater when shapes had an "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 happen 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 person, with a "self" inside the shape reduced the reaction times as compared with when a "other" inside the shape ($mean_{diff}$ = -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

[Summary of results] 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. We found robust effect that good character was prioritized in the shape-label matching task, regardless how good character was represented (single word or behavioral description). Moreover, the prioritization of good character was not driven by valence itself, i.e., "good" vs "bad". Instead, this effect was modulated by a self-referential processing: prioritization only occurred when moral characters are self-referential (experiments 3a, 3b, and 6b). Finally, the prioritization of good character was modulated by self-referential information even when either the self- or character-related information was irrelevant to experimental task (experiment 4a and 4b). In

contrast, for other referential condition, explicitly or implicitly, the performance of good character was worse than neutral character. Together, these results highlighted the importance of the self in perceiving more character information, suggested a spontaneous self-referential process when moral character is involved in perceptual decision-making. These results contribute to a growing literature on the social and relational nature of perception [Xiao, Coppin, and Bavel (2016); Freeman, Stolier, and Brooks (2020); hafri_perception_2021].

[Effect of good character] The evidence for the effect of moral character on perceptual decision-making is robust in our study. Previous studies reported the effect of morality on perception but the results and the mechanisms were disputed. For example, (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 studies 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 caused by semantic priming instead of morality (Firestone & Scholl, 2015; Jussim et al., 2016). In the current study, the associative learning task allowed us to eliminate the semantic priming. First, only a few pairs of stimuli was used and different stimuli were different in many dimensions, makes it impossible for priming between them. Second, stimuli that used to represent moral character are neutral stimuli before the associative learning. Moreover, in experiment 1c where participants first associate moral behaviors with neutral names and then paired names with neutral shapes, we still found that effect of moral character, suggesting that it is the moral content instead of other features of labels/names that drives the prioritization effect. Finally, consistent effect across all eight different samples in the first set of experiments further confirmed that the prioritization effect of good character found in our paradigm is robust.

Previous moral perception studies usually reported a negativity effect, i.e., information related to bad moral character are processed better (Anderson et al., 2011;

Eiserbeck & Abdel Rahman, 2020). For instance, Anderson et al. (2011) reported the faces associated with negative social behaviors dominated the awareness for longer time than those associated with neutral or positive behaviors. This discrepancy between previous results and the current study may resulted from differences in the task: while in many previous moral perception studies, the participants were asked to detect the existence of a stimuli, the current task asked participants to recognize a pattern. In other words, previous studies targeted early stages of perception while the current task focus more on the decision-making at relative later stage of information processing. This discrepancy is consistent with the pattern found in studies with emotional stimuli (Pool, Brosch, Delplanque, & Sander, 2016).

[Self-binding as a novel explanation] We expanded previous moral perception studies by examining two possible explanations of the prioritization of good character: valence effect or self-referential process. Our results revealed that prioritization of good character is modulated by self-relatedness of the character information: when the good character was prioritized when it was related to self, even when the self-relatedness was task irrelevant. By contrast, when good character information was no longer prioritized when it was associated with non-self. The modulation effect of self-referential process was large when the relationship between moral character and the self was explicit. More importantly, the effect persisted when the relationship between moral character and the self information was implicit, suggesting a spontaneous self-referential when both information were presented. An possible explanation for this spontaneous self-referential of good character is that moral-self is central to our identity (Freitas, Cikara, Grossmann, & Schlegel, 2017; Strohminger, Knobe, & Newman, 2017) and the motivation to maintain a moral-self view also influenced the perceptual decision-making.

[Beyond the debate about penetration of perception] Although the results here revealed prioritization of good character in perceptual decision-making, we did not claim that the moral-self motivation *penetrates* perception. Perceptual decision-making process

include processes more than just encoding the sensory inputs, we need more computational models that can account the nuance of behavioral data and/or related data collected from other modules. For example, sequential sampling models suggest that, when making a perceptual decision, the agent is continuously accumulate 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 model to perceptual matching tasks also suggest that different processes may contributed 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 need to disentangle the cognitive processes underlying the prioritization of good character.

References

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

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

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. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, e229. https://doi.org/10.1017/S0140525X15000965

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 [Journal Article]. *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? [Journal Article]. *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 Article]. *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

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. (2019). Reaction times and other skewed distributions: Problems with the mean and the median [Preprint]. *Meta-Psychology*. https://doi.org/10.1101/383935

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

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., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching [Journal Article]. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1105–1117. https://doi.org/10.1037/a0029792

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

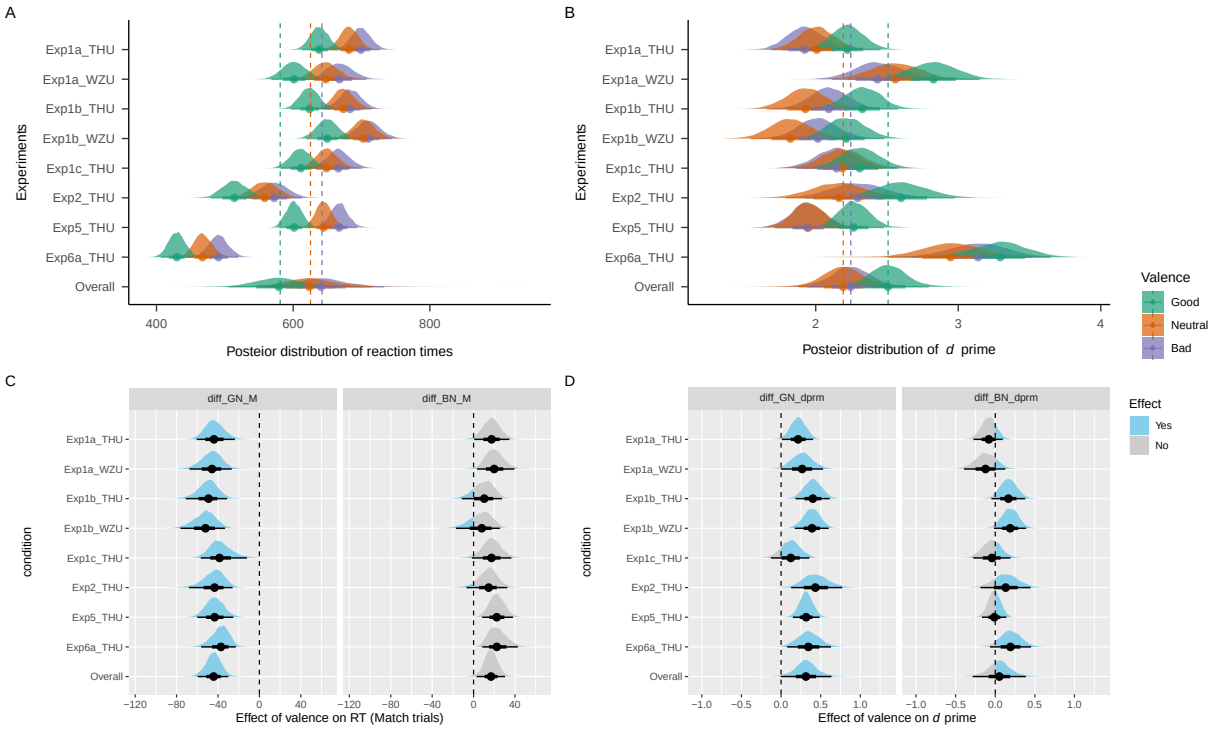

Figure 1. Effect of moral character on RT and d'

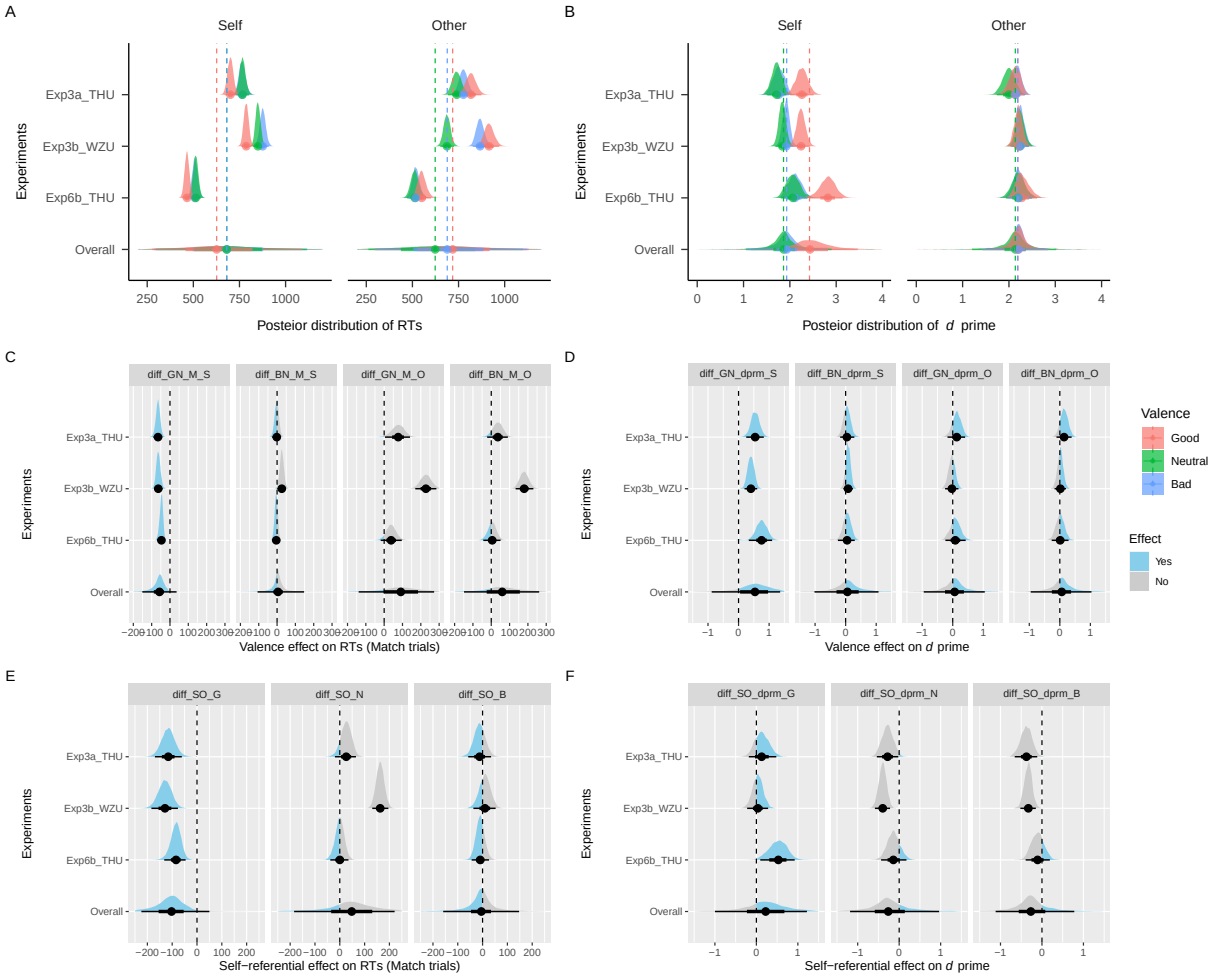

Figure 2. Interaction between moral character and self-referential

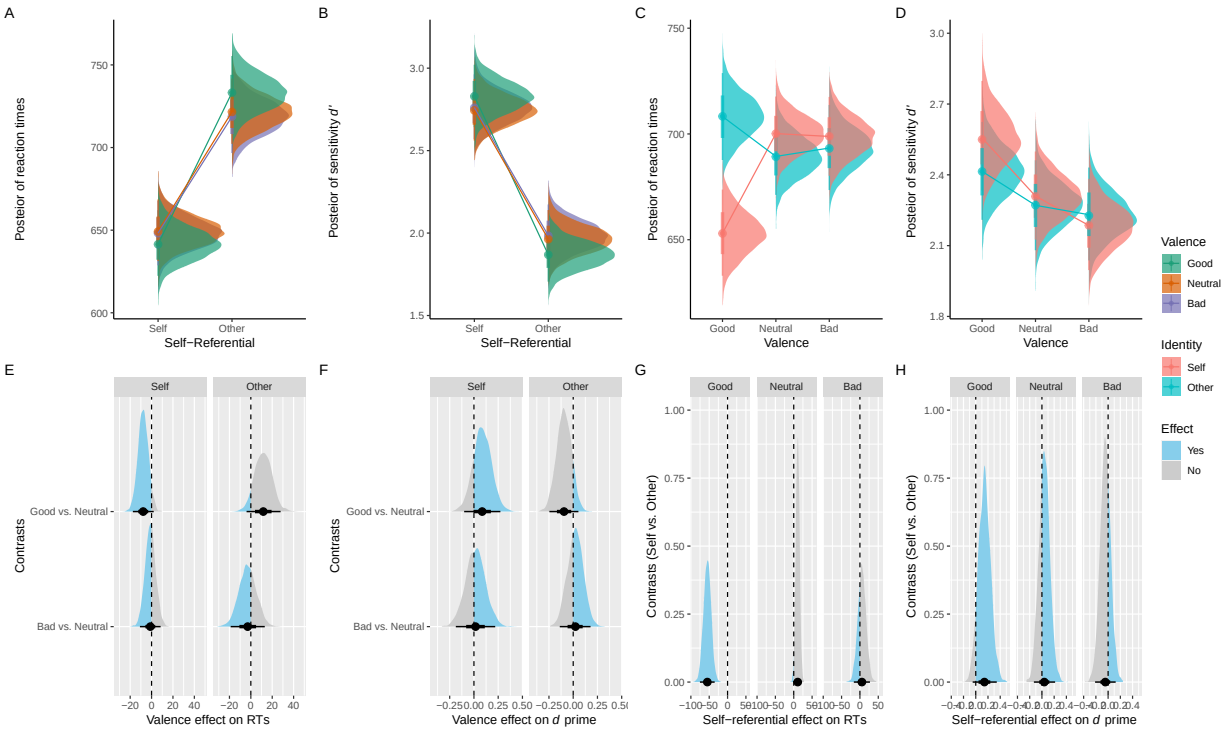

Figure 3. Experiment 4: Implicit binding between good character and the self.